

Data-Dependent Regularization

Adrian Corduneanu and Tommi Jaakkola
MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139

Abstract

Information regularization is a principle for assigning labels to unlabeled data points in a semi-supervised setting. The broader principle is based on finding labels that minimize the information induced between examples and labels relative to a topology over the examples; any label variation within a small local region of examples ties together the identities of examples and their labels. Such variation should be minimized unless supported directly or indirectly by the available labeled examples. The principle can be cast in terms of Tikhonov style regularization for maximizing likelihood of labeled examples with an information theoretic regularization penalty. We consider two ways of representing the topology over examples, either based on complete knowledge of the marginal density, or by grouping together examples whose labels should be related. We discuss the learning algorithms and sample complexity issues that result from each representation.

1 Introduction

A substantial number of algorithms and methods exist for solving supervised learning problems with little or no assumptions about the distribution generating the samples. Semi-supervised learning methods, in contrast, have to rely on assumptions about the problem so as to relate the available unlabeled data to possible class decisions. The most common such assumption is the cluster assumption (see Chapter 1, or Seeger [2001]) that, loosely speaking, prefers class decisions that cut between rather than through clusters of unlabeled points. The effect of the assumption is that it can significantly reduce the set of possible (reasonable) decisions that need to be considered in response to a few labeled examples. The same effect can be also achieved through representational constraints (e.g., Blum and Mitchell [1998]).

The definition of what constitutes a cluster and how the cluster assumption is formalized varies from one method to another. For example, clusters may be defined in terms of a weighted graph so that class decisions correspond to a graph partition

Szummer and Jaakkola [2001], Blum and Chawla [2001], A.Blum et al. [2004]. In a regularization setting, the graph may be used to introduce a smoothness penalty on the discriminant function so as to limit how the discriminant function can change within graph neighborhoods (e.g., see Chapter 11). Alternatively, we may define a model for each cluster via generative mixture models, and associate a single class decision (distribution over classes) with each mixture component (e.g., see Chapter 4).

The strength of the bias from unlabeled data can be directly controlled via the regularization parameter or by weighting likelihoods corresponding to labeled and unlabeled data. The choice of the weight may have a substantial effect on the resulting classifier, however (e.g., Corduneanu and Jaakkola [2002]).

We approach here the semi-supervised learning problem as a regularization problem, consistent with the broader cluster assumption, but define the regularization penalty by appealing to information theory. The key idea is to express the penalty as a bit cost of deviating decisions from those consistent with some assumed structure over the unlabeled examples. In our case the structure corresponds to a collection of overlapping sets or regions that play a role similar to clusters; decisions are biased to be the same within each set and their specification is tied to the marginal distribution over the examples. In practice, the sets can be derived from weighted graph neighborhoods for discrete objects or from ϵ -balls covering the unlabeled points.

We begin by introducing the overall *information regularization* principle. The structure of the remaining sections is modeled after Figure 1, successively elaborating the principle under variations in the example space, type of unlabeled data that is available, and which modeling assumptions we are willing to make.

Consider a typical semi-supervised learning problem with a few labeled examples $((x_1, y_1), \dots, (x_l, y_l))$ and a large number of unlabeled examples (x_{l+1}, \dots, x_n) or the marginal distribution $p(x)$. We assume that the labels are discrete taking values in $\mathcal{Y} = \{1, \dots, M\}$ for some finite M . The goal is to estimate the conditional distributions $Q(y|x)$ associated with each available example x (labeled or unlabeled).

We will introduce the information regularization approach here from two alternative perspectives: smoothness and communication. By smoothness we mean constraining how $Q(y|x)$ is allowed to vary from one point to another. The smoothness preference is expressed as a regularization penalty over different choices of $Q(\cdot|x)$, $x \in \mathcal{X}$. The communication perspective, on the other hand, characterizes the regularization penalty in terms of the cost of encoding labels for all the points using $Q(y|x)$ relative to a basic coding scheme.

In either case the key role is played by a collection of regions, denoted by \mathcal{R} . Each region $R \in \mathcal{R}$ represents a set of a priori equivalent examples. In other words, in the absence of any other information, we would prefer to associate the same distribution of labels with all $x \in R$. Figure 2 illustrates two possible overlapping regions. We will use these regions to exemplify the basic ideas.

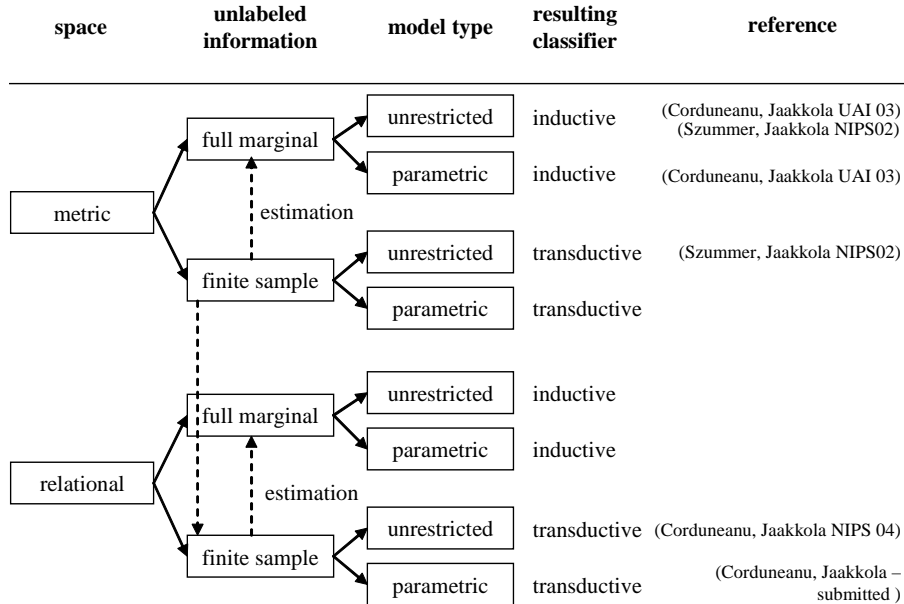


Figure 1: Outline of information regularization methods under different assumptions about the space, data, and model. Dotted arrows indicate that one setting can be cast as another through a simple transformation (estimation, or relations derived from metric)

1.1 Regions and smoothness

Consider the six unlabeled examples in region R in Figure 2. We assume that each point has the same probability of being a member of the region so that $P(x|R) = 1/6$. The membership probabilities provide an additional degree of freedom for specifying smoothness constraints. Given the region R and the membership probabilities $P(x|R)$, $x \in R$, we would like to introduce a penalty for any variation in the conditionals $Q(y|x)$ across the examples in the region. A natural choice for this penalty is the KL-divergence between each conditional $Q(y|x)$ and the best common choice $Q(y|R)$:

$$I_R(x; y) = \min_{Q(\cdot|R)} \sum_{x \in R} P(x|R) \sum_{y \in \mathcal{Y}} Q(y|x) \log \frac{Q(y|x)}{Q(y|R)} \quad (1)$$

$$= \sum_{x \in R} P(x|R) \sum_{y \in \mathcal{Y}} Q(y|x) \log \frac{Q(y|x)}{Q(y|R)} \quad (2)$$

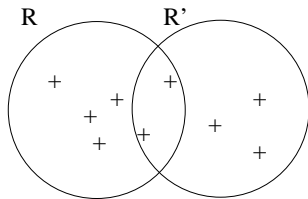


Figure 2: Example regions.

where $Q(y|R) = \sum_{x \in R} P(x|R)Q(y|x)$.¹ Note that we can interpret the result as the mutual information between x and y within the region so long as the joint distribution $Q(x, y)$ is defined as $Q(y|x)P(x|R)$. The mutual information involves no prior penalty on what the common distribution should be; $I_R(x, y)$ is zero if all the points in the region are labeled $y = 1$ or all of them have entirely uncertain conditionals $Q(y|x) = 1/M$.

Suppose now that some of the six examples in region R have been labeled. We will formulate the resulting estimation task as a regularization problem with the mutual information serving as a regularization penalty. To this end, let Q refer collectively to the parameters $Q(\cdot|x), x \in R$. Define $J(Q) = I_R(x; y)$ (which we will extend shortly to multiple regions) so that the penalized maximum likelihood criterion is given by

$$\sum_{i=1}^l \log Q(y_i|x_i) - \lambda J(Q)$$

where λ is a regularization parameter that balances the fit to the available labeled points and the smoothness bias expressed by $J(Q)$. If only one of the six points is labeled, all the points in the region will be labeled with the observed label. This is because the value of the regularizer is independent of the common choice within the region but biases any differences within the region. In case of two distinctly labeled points, the remaining points would be labeled such that the conditionals $Q(y|x)$ assign all their weight equally to the two observed labels while excluding all others. The conditionals associated with the labeled points would be drawn towards their respective labels, also excluding other than observed label values.

1.1.1 Multiple regions

In the single region case the labels for unlabeled points are pulled equally towards the optimized common distribution without further distinguishing between the points. The notion of locality arises from multiple regions, such as $\mathcal{R} = \{R, R'\}$ in the figure. In this setting, the overall regularization penalty must be a (weighted) average of the

¹ $I_R(x; y)$ is exactly the general Jensen-Shannon divergence between $Q(\cdot|x)$ for all $x \in R$, weighted by $P(x|R)$

individual region penalties:

$$J(Q) = \sum_{R \in \mathcal{R}} \gamma(R) I_R(x; y)$$

where $\gamma(R)$ represents the weight of region R , where the choice of $\gamma(R)$ is a modeling decision. $\gamma(R)$ expresses a priori belief in the relative importance of the regions, thus it is not necessarily related to $P(R) = \int_R p(x) dx$, the probability of region R derived from the generative distribution of the data.

In Figure 2 there are three sets of equivalent points that are not further distinguished in this regularizer. They are $R \setminus R'$, $R \cap R'$, and $R' \setminus R$. We call these sets that are not further partitioned by other regions *atomic regions*. By introducing more regions, we partition the space into smaller atomic regions and thus can make finer distinctions between the conditional distributions associated with the points; within each atomic region, the conditional distributions can differ only if some of the points are explicitly labeled.

A sequence of overlapping regions can mediate influence between the conditionals associated with more “remote” points, those that do not appear in a common region. For example, labeling any point in $R \setminus R'$ will also set all the labels in $R' \setminus R$ via the intersection. Note, however, that labeling the points in the intersection would not completely remove this influence; the Markov properties associated with the regions pertain to the conditional distributions, not labels directly.

The choice of the regions, region weights $\gamma(R)$, and the membership probabilities $P(x|R)$ will change the regularizer. While these provide additional degrees of freedom that have to be set (or learned), there are nevertheless simple ways of specifying them directly based on the problem. For example, suppose we are given a weighted undirected graph with vertex set V , edge set E , and edge weights $w(u, v)$ associated with any $(u, v) \in E$. Then we can simply associate the regions with edges, specify equal membership probabilities for vertices in each edge, and set $\gamma(R)$ equal to the weight of the corresponding edge in the graph. The resulting regularizer is analogous to the graph based regularizers for discriminant functions except that it is cast in terms of conditional probabilities.

1.2 Communication principle

The information regularization objective can be also derived from a communication principle. Suppose we have the same collection of regions \mathcal{R} , region weights $\gamma(R)$, membership probabilities $P(x|R)$, and the conditionals $\{Q(y|x)\}$ associated with the points. The regularizer is defined as the bit rate of communicating labels for points according to the following communication game. In this scheme, the regions, points, and labels are sampled as follows. First, we select a region $R \in \mathcal{R}$ with probability (proportional to) $\gamma(R)$, then a point within the region according to the membership probabilities $P(x|R)$, and finally the label y from $Q(y|x)$. The label is then communicated to the receiver using a coding scheme tailored to the region, i.e., on the basis of

$Q(y|R)$. The receiver is assumed to have prior access to x , R , and the region specific coding scheme. Under these assumptions, the amount of information that must be sent to the receiver to accurately reconstruct the samples on average is

$$J(Q) = \sum_{R \in \mathcal{R}} \gamma(R) I_R(x; y)$$

which is the regularizer previously defined. Equivalently, we can rewrite the regularizer as:

$$J(Q) = I(x; y) - I(R; y)$$

Therefore the communication principle aims to minimize any information x has to communicate about y beyond what has already been communicated by the region from which x was drawn. This information is minimal when the label within each region does not depend on which x we sampled.

2 Information regularization on metric spaces

We adapt here the information regularization principle to the setting where \mathcal{X} is a metric space and assume that its metric is correlated with the labeling of points. In other words, points that are close according to the metric are likely to have the same label. For example, if \mathcal{X} is a real vector space the metric could be the Euclidean distance between the points, possibly weighted by feature relevance. Using a metric to introduce a bias in semi-supervised learning is quite common, and many existing algorithms require an explicit or implicit metric.

2.1 Full knowledge of the marginal

We begin by considering the ideal situation in which we have access to unlimited unlabeled data, which, together with the metric, amounts to knowing the marginal density $p(x)$. In this case the information regularizer will relate the structure of $p(x)$ to the possible labelings of points. While we develop the ideas in the context of knowing the marginal, the resulting algorithms apply also to finite sample cases, by replacing $p(x)$ with an empirical estimate.

2.1.1 The information regularizer

In order to construct the regularizer we need to specify how the regions cover the metric space along with the weights $\gamma(R)$ associated with the regions. The cover \mathcal{R} should provide connected and significantly overlapping regions. This is necessary since labeling one point can only affect another if they can be connected through a path of overlapping regions.

In covering the space we have to balance the size of the regions with their overlap. We derive here the form of the regularizer in the limit of vanishing but highly overlapping regions. Under mild constraints about how the limit is taken, the resulting regularizer is the same. The limiting form has the additional benefit that it no longer requires us to engineer a particular covering of the space.

We choose the regions such that as their size approaches 0, the overlap between neighbors approaches 100% (this is required for smoothness). In the limit therefore each point belongs to infinitely many regions, resulting in an infinite sum of local regularizers. An appropriate choice of λ , the regularization parameter, is needed to re-scale the regularizer to take into account this increase.

In choosing the cover \mathcal{R} care must be taken not to introduce systematic biases into the regularizer. Assuming that \mathcal{X} has vector space structure, we can cover it with a homogeneous set of overlapping regions of identical shape: regions centered at the axis-parallel lattice points spaced at distance l' . In what follows the regions are going to be axis-parallel cubes of length l , where l is much larger than l' . Because \mathcal{R} covers \mathcal{X} uniformly, we can weight the regions based on the marginal density, i.e., $\gamma(R) = P(R)$ up to a multiplicative constant.

Assuming that l and l' are such that l/l' is an integer, each (non-lattice) point belongs to $(l/l')^d$ cubic regions, where d is the dimension of the vector space. Let \mathcal{R}' be the *partitioning* of \mathcal{R} into atomic lattice cubes of length l' . Each region in \mathcal{R} is partitioned into $(l/l')^d$ disjoint atomic cubes from \mathcal{R}' , and each atomic cube is contained in $(l/l')^d$ overlapping regions from \mathcal{R} . We may now rewrite the global regularizer as a sum over the partition \mathcal{R}' :

$$\begin{aligned} J(p) &= \lim_{l \rightarrow 0} \sum_{R \in \mathcal{R}} P(R) I_R(x; y) = \lim_{l \rightarrow 0} \sum_{R' \in \mathcal{R}'} P(R') \sum_{R \supseteq R'} I_R(x; y) = \\ &= (l/l')^d \lim_{l' \rightarrow 0} \sum_{R' \in \mathcal{R}'} P(R') I_{R'}(x; y) = \lim_{l' \rightarrow 0} (l/l')^d \cdot \int_{\mathcal{X}} p(x) \frac{dI_R(x; y)}{dx} dx \end{aligned}$$

Note that the factor in front of the integral can be factored into the regularization parameter λ as a multiplicative constant.

Infinitesimal Mutual Information

We derive the local mutual information as the diameter of R approaches 0. If x_0 is the expectation of x over R , mutual information takes the following asymptotic form:

$$I_R(x; y) = \frac{1}{2} \text{tr} (\mathbf{Var}_R [x] F(x_0)) + \mathcal{O}(\text{diam}(R)^3)$$

where $F(x) = \mathbf{E}_{Q(y|x)} [\nabla_x \log Q(y|x) \cdot \nabla_x \log Q(y|x)^\top]$ is the Fisher information and $\mathbf{Var}_R [x]$ is the covariance of $p_R(x)$ (for a proof of this result see Corduneanu and Jaakkola [2003]). Note that since the covariance is $\mathcal{O}(\text{diam}(R)^2)$, $I_R(x; y) \rightarrow 0$ as $\text{diam}(R) \rightarrow 0$. Therefore $\lim_{\text{diam}(R) \rightarrow 0} I_R(x; y) / \text{diam}(R)^2$ is well-defined, and this is

the infinitesimal quantity that we will integrate to obtain $J(p)^2$:

$$J(p) = \int_{\mathcal{X}} p(x) \mathbf{tr} \left(F(x) \lim_{\text{diam}(R) \rightarrow 0} \frac{\mathbf{Var}_R[x]}{\text{diam}(R)^2} \right) dx$$

Given this form of the regularizer we can argue that regions in the shape of a cube are indeed appropriate. We start from the principle that the regularizer should not introduce any systematic directional bias in penalizing changes in the label. If the diameter of a region R is small enough, $p_R(x)$ is almost uniform, and $p(y = 1|x)$ can be approximated well by $\mathbf{v} \cdot x + c$, where \mathbf{v} is the direction of highest variation. In this setting we have the following result (Corduneanu and Jaakkola [2003]):

Theorem 1 *Let R be such that $\text{diam}(R) = 1$. The local information regularizer is independent of $\mathbf{v}/\|\mathbf{v}\|$ if and only if $\mathbf{Var}_R[\cdot]$ is a multiple of the identity.*

Proof We have $F(x_0) = \mathbf{v}\mathbf{v}^\top$. The relevant quantity that should be independent of $\mathbf{v}/\|\mathbf{v}\|$ is therefore $\mathbf{v}^\top \mathbf{Var}_R[\cdot] \mathbf{v}$. Let $v = \Phi_i/\|\Phi_i\|$, where Φ_i is an eigenvector of $\mathbf{Var}_R[\cdot]$ of eigenvalue ϕ_i . Then $\mathbf{v}^\top \mathbf{Var}_R[\cdot] \mathbf{v} = \phi_i$ should not depend on the eigenvector. It follows that $\mathbf{Var}_R[\cdot]$ has equal eigenvalues, thus $\mathbf{Var}_R[\cdot] = \phi \mathbf{I}$. The converse is trivial.

It follows that in order to remove any directional bias, $\mathbf{Var}_R[x] \approx \text{diam}(R)^2 \cdot \mathbf{I}$, as it is the case if R is a cube or a sphere. We thus reach our final form of the information regularizer for metric space when the marginal is fully known:

$$J(p) = \int_{\mathcal{X}} p(x) \mathbf{tr}(F(x)) dx \tag{3}$$

Note the the dependence of \mathcal{R} is only implicit.

2.1.2 Classification algorithm

We would like to estimate a label confidence $Q(\cdot|x)$ (that is, a *soft* label in $[0, 1]^M$) for every $x \in \mathcal{X}$ given the knowledge of $p(x)$, and a labeled sample $\{(x_i, y_i)\}_{i=1 \dots l}$. The information regularization principle requires us to maximize the regularized log-likelihood:

$$\max_{\{Q(y|x); x \in \mathcal{X}, y \in \mathcal{Y}\}} \sum_{i=1}^l \log Q(y_i|x_i) - \lambda \int_{\mathcal{X}} p(x) \mathbf{tr}(F(x)) dx \tag{4}$$

where $F(x) = \mathbf{E}_{Q(y|x)} [\nabla_x \log Q(y|x) \cdot \nabla_x \log Q(y|x)^\top]$, and the maximization is subject to $0 \leq Q(y|x) \leq 1$ and $\sum_{y \in \mathcal{Y}} Q(y|x) = 1$.

²To be consistent with the derivation of $J(p)$, we should normalize $I_R(x; y)$ by $\text{diam}(R)^d$, but unless $d = 2$ the regularizer would be either 0 or ∞ . We can afford to choose the convenient normalization without compromising the principle because we are free to choose λ

It is interesting that the above optimization defines a labeling even in a completely unrestricted non-parametric setting (save for differentiability constraints on $Q(\cdot|x)$). In this situation labels of distinct data points are related only through the information regularizer. We show that if we fix the values of the labels at the observed labeled samples, $Q(y_i|x_i) = P_0(y_i|x_i)$, for all $i = 1 \dots l$, the regularizer extends $Q(y|x)$ to unobserved x 's uniquely. In what follows, we restrict the analysis to binary classification ($\mathcal{Y} = \{-1, 1\}$).

We cast the optimization as solving a differential equation that characterizes the optimal conditional. The conditional that minimizes the regularizer $\int p(x) \mathbf{tr}(F(x))$ is a differentiable function (except maybe at the labeled samples, where it is only continuous) that satisfies the Euler-Lagrange condition:

$$\nabla_x \log p(x) \nabla_x Q(1|x)^\top + \mathbf{tr}(\nabla_{xx}^2 Q(1|x)) + \frac{1}{2} \frac{Q(1|x) - Q(-1|x)}{Q(1|x)Q(-1|x)} \|\nabla_x Q(1|x)\|^2 = 0$$

(Corduneanu and Jaakkola [2003])

This differential equation defines a unique solution given the natural boundary conditions $p(x) = 0$ and $\nabla_x Q(y|x) = 0$ at infinity, as well as the labels $P_0(y_i|x_i)$ at labeled samples.

In order to optimize (4) one could solve the differential equation for various values $\{P_0(y_i|x_i)\}_{i=1 \dots l}$, then optimize with respect to $P_0(y_i|x_i)$. Unfortunately, solving the differential equation numerically involves discretizing \mathcal{X} , which is impractical for all but low dimensional spaces. That is why the non-parametric but inductive (find a label for each point in \mathcal{X}) information regularization is of more theoretical than practical interest.

Nevertheless, if \mathcal{X} is the one-dimensional real line the differential equation can be solved analytically (Corduneanu and Jaakkola [2003]). We present the solution here to illustrate the type of biases imposed by the information regularizer. When \mathcal{X} is one dimensional, the labeled samples x_1, x_2, \dots, x_l split the real line into disjoint intervals; thus if $P_0(y|x_i)$ are given, the differential equation can be solved independently on each interval determined by the samples. The solution only depends on the labels of the endpoints, and is given by the following:

$$Q(1|x) = \frac{1}{1 + \tan^2\left(-c \int \frac{1}{p(x)}\right)}$$

where c and the additive constant in $\int 1/p$ can be determined from the values of the conditional at the endpoints. These two parameters need not be the same on different intervals.

Figure 3 shows the influence of various $p(x)$ on $Q(1|x)$ through information regularization under the boundary conditions $P(y = 1|x = 0) = 0.9$ and $P(y = 1|x = 1) = 0.1$. The property of preferring changes in the label in regions of low data density is evident. Note that the optimal $P(y|x)$ will always be between its values at the boundary; otherwise for some $x_1 \neq x_2$ we would have $P(y|x_1) = P(y|x_2)$,

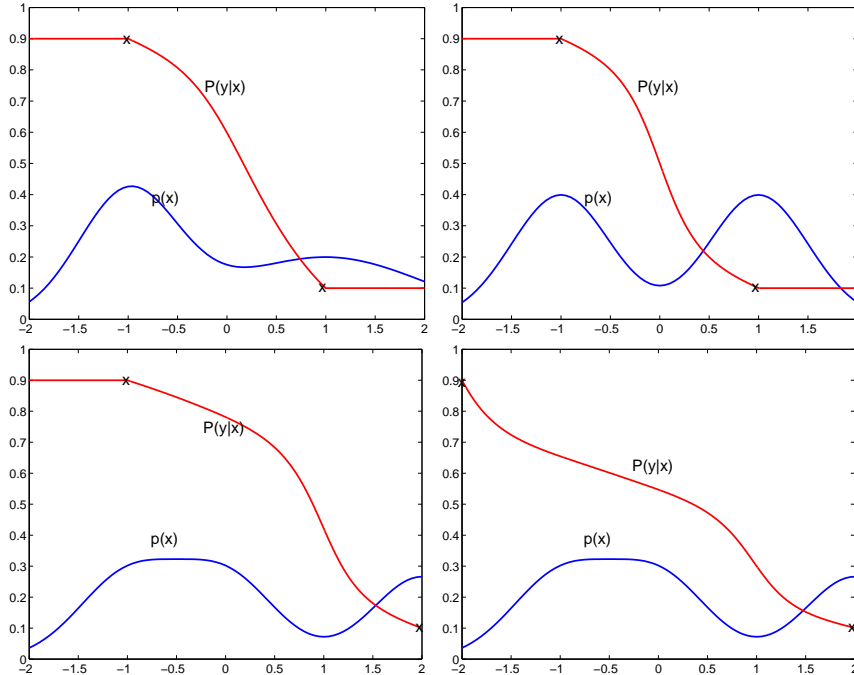


Figure 3: Non-parametric conditionals that minimize the information regularizer for various one-dimensional data densities while the label at boundary labeled points is fixed

and because the cumulative variation is minimized, necessarily $P(y|x) = P(y|x_1)$ for every $x \in [x_1, x_2]$.

2.1.3 Learning theoretic properties

We extend the analysis of information regularization on metric spaces given the full knowledge of the marginal with a learning theoretical framework. The aim is to show that the information regularizer captures the learning complexity, in the sense that bounding it makes the labels learnable without any additional assumptions about $\{Q(y|x)\}_{x \in \mathcal{X}, y \in \mathcal{Y}}$. Because the setting is non-parametric, and the only link that relates labels of distinct points is the information regularizer, $\{Q(y|x)\}_{x \in \mathcal{X}, y \in \mathcal{Y}}$ would not be learnable without placing a constraint on the information regularizer. While the learning framework is general, due to technical constraints³ we derive an explicit sample-size bound only for binary classification when \mathcal{X} is one-dimensional.

We need to formalize the concepts, the concept class (from which to learn them), and a measure of achievement consistent with (4). The key is then to show that the task is learnable in terms of the complexity of the concept class.

Standard PAC-learning of indicator functions of class membership will not suffice for our purpose. Indeed, conditionals with very small information regularizer can

³Only in one dimension the labeled points give rise to segments that can be optimized independently.

still have very complex decision boundaries, of infinite VC-dimension. Instead, we rely on the *p-concept* (Kearns and Schapire [1994]) model of learning full conditional densities: concepts are functions $Q(y|x) : \mathcal{X} \rightarrow [0, 1]$. Then the concept class is that of conditionals with bounded information regularizer:

$$\mathcal{J}_\gamma(p) = \left\{ Q : \int_{\mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} Q(y|x) \|\nabla_x \log Q(y|x)\|^2 dx \leq \gamma \right\}$$

We measure the quality of learning by a loss function $L_Q : \mathcal{X} \times \mathcal{Y} \rightarrow [0, \infty)$. This can be the log-loss $-\log Q(y|x)$ associated with maximizing likelihood, or the square loss $(Q(y|x) - 1)^2$. The goal is to estimate from a labeled sample a concept Q_{opt} from $\mathcal{J}_\gamma(p)$ that minimizes the expected loss $\mathbf{E}_{p(x)P(y|x)} [L_Q]$, where $P(y|x)$ is the true conditional.

One cannot devise an algorithm that optimizes the expected loss directly, because this quantity depends on the unknown $P(y|x)$. We make the standard approximation of estimating Q_{opt} by minimizing instead the empirical estimate of the expected loss from the labeled sample:

$$\hat{Q} = \arg \min_{Q \in \mathcal{J}_\gamma(p)} \hat{\mathbf{E}} [L_Q] = \arg \min_Q \frac{1}{l} \sum_{i=1}^l L_Q(x_i, y_i)$$

If the loss function is the log-loss, finding \hat{Q} is equivalent to maximizing the information regularization objective (4) for a specific value of λ . However, we will present the learning bound for the square loss, as it is bounded and easier to work with. A similar result holds for the log-loss by using the equivalence results between the log-loss and square-loss presented in (Abe et al. [2001]).

The question is how different \hat{Q} (estimated from the sample) and Q_{opt} (estimated from the true conditional) can be due to this approximation. Learning theoretical results provide guarantees that given enough labeled samples the minimization of $\hat{\mathbf{E}} [L_Q]$ and $\mathbf{E}_{p(x)P(y|x)} [L_Q]$ are equivalent. We say the task is learnable if with high probability in the sample the empirical loss converges to the true loss uniformly for all concepts as $l \rightarrow \infty$. This guarantees that $\mathbf{E} [L_{\hat{Q}}]$ approximates $\mathbf{E} [L_{Q_{opt}}]$ well. Formally,

$$\mathbb{P}\{\exists Q \in \mathcal{J}_\gamma(p) : |\hat{\mathbf{E}} [L_Q] - \mathbf{E} [L_Q]| > \epsilon\} \leq \delta \tag{5}$$

where the probability is with respect to all samples of size l . The inequality should hold for l polynomially large in $1/\epsilon, 1/\delta, 1/\gamma$.

We have the following sample complexity bound on the square loss, derived in Corduneanu and Jaakkola [2003]:

Theorem 2 *Let $\epsilon, \delta > 0$. Then*

$$\mathbb{P}\{\exists Q \in \mathcal{J}_\gamma(p) : |\hat{\mathbf{E}} [L_Q] - \mathbf{E} [L_Q]| > \epsilon\} < \delta$$

where the probability is over samples of size l greater than

$$\mathcal{O}\left(\frac{1}{\epsilon^4}\left(\log\frac{1}{\epsilon}\right)\left[\log\frac{1}{\delta}+c_p(m_p^{-1}(\epsilon^2))+\frac{\gamma}{(m_p^{-1}(\epsilon^2))^2}\right]\right)$$

Here $m_p(\alpha) = \mathbb{P}\{x : p(x) \leq \alpha\}$, and $c_p(\alpha)$ is the number of disconnected sets in $\{x : p(x) > \alpha\}$.

The quantities $m_p(\cdot)$ and $c_p(\cdot)$ characterize how difficult the classification is due to the structure of $p(x)$. Learning is more difficult when significant probability mass lies in regions of small $p(x)$ because in such regions the variation of $Q(y|x)$ is less constrained. Also, the larger $c_p(\cdot)$ is, the labels of more “clusters” need to be learned from labeled data. The two measures of complexity are well-behaved for the useful densities. Densities of bounded support, Laplace and Gaussian, as well mixtures of these have $m_p(\alpha) < u\alpha$, where u is some constant. Mixtures of single-mode densities have $c_p(\alpha)$ bounded by the number of mixtures.

2.2 Finite unlabeled sample

We discuss here classification by information regularization when \mathcal{X} is endowed with a metric but the true marginal $p(x)$ is unknown save a large unlabeled sample (x_{l+1}, \dots, x_n) . In practice we might already have a domain specific model (class) of how the labels are generated and we show how to apply information regularization if the labels must come from a parametric family $Q(y|x, \theta)$.

Although it is possible to approach this scenario directly by partitioning the space into regions as in Szummer and Jaakkola [2002], here we reduce the task to the situation in which the full marginal is known by replacing the full marginal with an empirical estimate obtained from the unlabeled sample.

We illustrate this method on logistic regression, in which we restrict the conditional to linear decision boundaries with the following parametric form: $Q(y|x; \theta) = \sigma(y\theta^\top x)$, where $y \in \{-1, 1\}$ and $\sigma(x) = 1/(1 + \exp(-x))$. The Fisher information is therefore $F(x; \theta) = \sigma(\theta^\top x)\sigma(-\theta^\top x)\theta\theta^\top$ and according to Equation 3 the information regularizer takes the form

$$\|\theta\|^2 \int \hat{p}(x)\sigma(\theta^\top x)\sigma(-\theta^\top x)dx$$

Here $\hat{p}(x)$ is the empirical estimate of the true marginal. We compare two ways of estimating $p(x)$, the empirical approximation $\frac{1}{n} \sum_{j=1}^n \delta(x - x'_j)$, as well as a Gaussian kernel density estimator. The empirical approximation leads to optimizing the following criterion:

$$\max_{\theta} \sum_{i=1}^l \log \sigma(y_i \theta^\top x_i) - \|\theta\|^2 \frac{\lambda}{n} \sum_{j=1}^n \sigma(\theta^\top x_j)\sigma(-\theta^\top x_j)$$

It is instructive to contrast this information regularization objective with the criterion optimized by *transductive SVM's*, as in Chapter 6. Changing the SVM loss function to logistic loss, transductive SVM/logistic regression optimizes:

$$\max_{\theta, y_1, \dots, y_n} \sum_{i=1}^n \log \sigma(y_i \theta^\top x_i) - \frac{\lambda}{2} \|\theta\|^2$$

over all labelings of unlabeled data. In contrast, our algorithm contains the unlabeled information in the regularizer.

The presented information regularization criterion can be easily optimized by gradient-ascent or Newton type algorithms. Note that the term $\sigma(\theta^\top x)\sigma(-\theta^\top x) = Q(1|x)Q(-1|x)$ focuses on the decision boundary. Therefore compared to the standard logistic regression regularizer $\|\theta\|^2$, we penalize more decision boundaries crossing regions of high data density. Also, the term makes the regularizer non-convex, making optimization potentially more difficult. This level of complexity is however unavoidable by any semi-supervised algorithm for logistic regression, because the structure of the problem introduces locally optimal decision boundaries.

If unlabeled data is limited, we may prefer a kernel estimate $\hat{p}(x) = \frac{1}{n} \sum_{j=1}^n K(x, x_j')$ to the empirical approximation, provided the regularization integral remains tractable. In logistic regression, if the kernels are Gaussian we can make the integral tractable by approximating $\sigma(\theta^\top x)\sigma(-\theta^\top x)$ with a degenerate Gaussian. Either from the Laplace approximation, or the Taylor expansion $\log(1 + e^x) \approx \log 2 + x/2 + x^2/8$, we derive the following approximation, as in Corduneanu and Jaakkola [2003]:

$$\sigma(\theta^\top x)\sigma(-\theta^\top x) \approx \frac{1}{4} \exp\left(-\frac{1}{4}(\theta^\top x)^2\right)$$

With this approximation computing the integral of the regularizer over the kernel centered μ of variance $\tau\mathbf{I}$ becomes integration of a Gaussian:

$$\begin{aligned} \frac{1}{4} \exp\left(-\frac{1}{4}(\theta^\top x)^2\right) \mathcal{N}(x; \mu, \tau\mathbf{I}) = \\ \frac{1}{4} \sqrt{\frac{\det \Sigma_\theta}{\det \tau\mathbf{I}}} \exp\left(-\frac{\mu^\top (\tau\mathbf{I} - \Sigma_\theta) \mu}{2\tau^2}\right) \mathcal{N}\left(x; \frac{\Sigma_\theta \mu}{\tau}, \Sigma_\theta\right) \end{aligned}$$

where $\Sigma_\theta = \left(\frac{1}{\tau}\mathbf{I} + \frac{1}{2}\theta\theta^\top\right)^{-1} = \tau \left[\mathbf{I} - \frac{1}{2}\theta\theta^\top / \left(\frac{1}{\tau} + \frac{1}{2}\|\theta\|^2\right)\right]$

After integration only the multiplicative factor remains:

$$\frac{1}{4} \left(1 + \frac{\tau}{2} \|\theta\|^2\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{4} \frac{(\theta^\top \mu)^2}{1 + \frac{\tau}{2} \|\theta\|^2}\right)$$

Therefore if we place a Gaussian kernel of variance $\tau\mathbf{I}$ at each sample x_j we obtain the following approximation to the information regularization penalty:

$$\frac{\|\theta\|^2}{\sqrt{1 + \frac{\tau}{2} \|\theta\|^2}} \frac{1}{4n} \sum_{j=1}^n \exp\left(-\frac{1}{4} \frac{(\theta^\top x_j)^2}{1 + \frac{\tau}{2} \|\theta\|^2}\right)$$

This regularizer can be also optimized by gradient ascent or Newton’s method.

2.2.1 Logistic regression experiments

We demonstrate the logistic information regularization algorithm as derived in the previous section on synthetic classification tasks. The data is generated from two bivariate Gaussian densities of equal covariance, a model in which the linear decision boundary can be Bayes optimal. However, the small number of labeled samples is not enough to accurately estimate the model, and we show that information regularization with unlabeled data can significantly improve error rates.

We compare a few criteria: logistic regression trained only on labeled data and regularized with the standard $\|\theta\|^2$; logistic regression regularized with the information regularizer derived from the empirical estimate to $p(x)$; and logistic regression with the information regularizer derived from a Gaussian kernel estimate of $p(x)$.

We have optimized the regularized likelihood $L(\theta)$ both with gradient ascent $\theta \leftarrow \theta + \alpha \nabla_{\theta} L(\theta)$, and with Newton’s method (iterative re-weighted least squares) $\theta \leftarrow \theta - \alpha \nabla_{\theta\theta}^2 L(\theta)^{-1} \nabla_{\theta} L(\theta)$ with similar results. Newton’s method converges with fewer iterations, but computing the Hessian becomes prohibitive if data is high dimensional, and convergence depends on stronger assumptions than those for gradient ascent. Gradient ascent is safer but slower.

We ran 100 experiments with data drawn from the same model and averaged the error rates to obtain statistically significant results. In Figure 4 (Corduneanu and Jaakkola [2003]) we have obtained the error rates on 5 labeled and 100 unlabeled samples. On each data set we initialized the iteration randomly multiple times. We set the kernel width τ of the Gaussian kernel approximation to the regularizer by standard cross-validation for density estimation. Nevertheless, on such large number of unlabeled samples the information regularizers derived from kernel and empirical estimates perform indistinguishable. They both outperform the standard supervised regularization significantly.

3 Information regularization and relational data

In a large number of classification domains we do not have a natural metric relevant to the classification task (correlating $Q(y|x)$ and $Q(y|x')$ for $x \neq x'$). In the absence of a metric biases about labelings are often naturally expressed in relational form. For example, consider the task of categorization of web pages in the presence of information about their link structure. It is natural to believe that pages that are linked in the same manner (common parents and common children) are biased to have similar topics even before we see any information about their content. Similarly, all other things being equal, pages that share common words are likely to have similar topics. In classifying gene function, genes whose protein products interact are more likely to participate in the same process with similar function; or in retrieving science

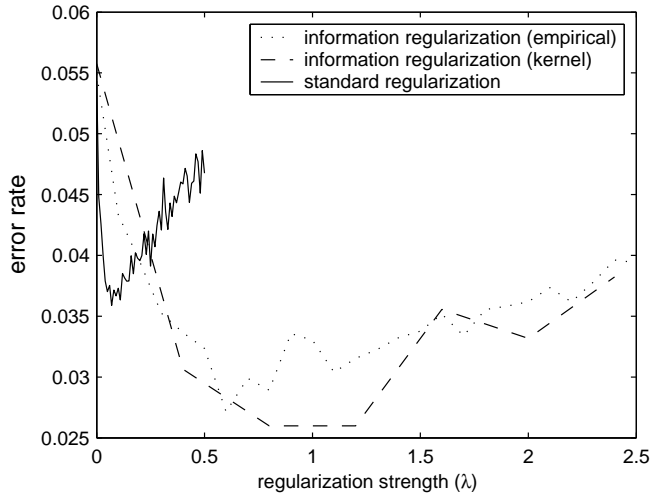


Figure 4: Average error rates of logistic regression with and without information regularization on 100 random selections of 5 labeled and 100 unlabeled samples from bivariate gaussian classes

publications, co-cited articles, or articles published in the same journal, are likely to have similar relevance assessments.

Relational classification is not new – it has been studied extensively from a Bayesian network perspective, as in Taskar et al. [2002]. Nevertheless, information regularization can exploit the relational structure with minimal assumptions about the distribution of data, even in a non-parametric, purely transductive context.

Let us begin by representing the relational constraints as a collection of regions (sets) \mathcal{R} , derived from observed examples (x_1, x_2, \dots, x_n) , where we expect the labels to be similar within each region. The regions here differ from the continuous case in that they are discrete subsets of indices $\{1, 2, \dots, n\}$ in the training set. It is useful to depict the region cover as a bipartite graph with points on one side and regions on the other, as in Figure 5. Note that regions can also be derived from a metric if such metric exists. For example, we could define regions centered at each observed data point of a certain radius. For this reason every algorithm discussed in this section is also applicable to finite sample metric settings.

We consider a generative process over the finite sample (x_1, x_2, \dots, x_n) by selecting a region R from \mathcal{R} with probability $\gamma(R)$, and then an observed point x_i from R according to the membership⁴ probability $P(i|R)$. The probabilities $\gamma(R)$ and $P(x|R)$ are task specific and must be selected such that $\sum_{R \in \mathcal{R}} \gamma(R)P(i|R) = P(x_i)$, the probability of sampling x_i from (x_1, \dots, x_n) . If the true marginal is known, then we can replace $P(x_i)$ with its true value; otherwise, a reasonable empirical estimate is $P(x_i) = 1/n$ for all $i = 1 \dots n$. If there is no reason to prefer one region over another, $\gamma(R)$ could be uniform on \mathcal{R} ; the constraint $P(x_i) = 1/n$ cannot be typically

⁴In the finite sample case we use the index of the example interchangeably with the example itself.

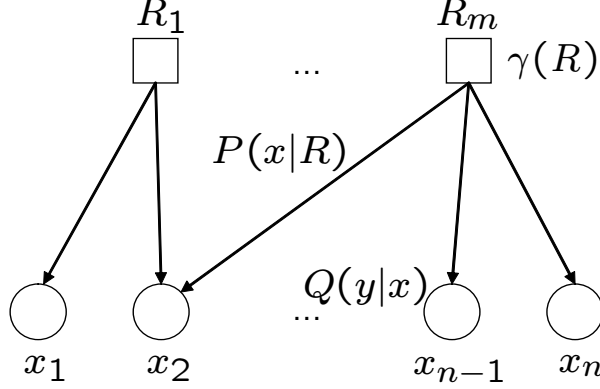


Figure 5: Covering of the observed samples with a set of relational regions represented as a bipartite graph. The lower nodes are the observed data points, and the upper nodes are the regions.

simultaneously enforced, however.

In this context the goal of classification is purely transductive: given the labels of the labeled training set, the classifier assigns labels to the unlabeled training set in a manner consistent with the relational biases \mathcal{R} . Nothing is inferred about unobserved $x \in \mathcal{X}$.

3.1 Non-parametric classification

Without constraining the family of label distributions $Q(y|x)$, the objective that must be optimized according the information regularization principle is:

$$\max_{\{Q(y|x_i)\}_{i=1\dots n}} \frac{1}{l} \sum_{i=1}^l \log Q(y_i|x_i) - \lambda J(Q; \mathcal{R})$$

where the information regularizer is given by

$$J(Q; \mathcal{R}) = \sum_{R \in \mathcal{R}} \gamma(R) I_R(x; y) = \sum_{R \in \mathcal{R}} \gamma(R) \sum_{j \in R} \sum_{y \in \mathcal{Y}} P(j|R) Q(y|x_j) \log \frac{Q(y|x_j)}{Q(y|R)}$$

where $Q(y|R) = \sum_{j \in R} P(j|R) Q(y|x_j)$ is the overall probability of y within the region.

As opposed to the continuous version of information regularization, the above objective depends on a finite set of parameters $\{Q(y|x_i)\}_{i=1\dots n}$, thus optimization is efficient. Moreover, in the non-parametric setting the objective is convex due to the

convexity of mutual information (Cover and Thomas [1991]). The following lemma from Corduneanu and Jaakkola [2004] formalizes the result:

Lemma 3 *The relational regularization objective for $\lambda > 0$ is a strictly convex function of the conditionals $\{Q(y|x_i)\}$ provided that 1) each point $i \in \{1, \dots, n\}$ belongs to at least one region containing at least two points, and 2) the membership probabilities $P(i|R)$ and $\gamma(R)$ are all non-zero.*

3.1.1 Distributed propagation algorithm

As in Corduneanu and Jaakkola [2004] we derive a local propagation algorithm for minimizing the relational regularization objective that is both easy to implement and provably convergent. The algorithm can be seen as a variant of the Blahut-Arimoto algorithm in rate-distortion theory (Blahut [1972]). We begin by rewriting each mutual information term $I_R(x; y)$ in the criterion

$$\begin{aligned} I_R(x; y) &= \sum_{j \in R} \sum_{y \in \mathcal{Y}} P(j|R) Q(y|x_j) \log \frac{Q(y|x_j)}{Q(y|R)} \\ &= \min_{Q_R(\cdot)} \sum_{j \in R} \sum_{y \in \mathcal{Y}} P(j|R) Q(y|x_j) \log \frac{Q(y|x_j)}{Q_R(y)} \end{aligned}$$

where the variational distribution $Q_R(y)$ can be chosen independently from $Q(y|x_j)$ but the unique minimum is attained when $Q_R(y) = Q(y|R) = \sum_{j \in R} P(j|R) Q(y|x_j)$. We can extend the regularizer over both $\{Q(y|x_i)\}$ and $\{Q_R(y)\}$ by defining

$$J(Q, Q_R; \mathcal{R}) = \sum_{R \in \mathcal{R}} \gamma(R) \sum_{j \in R} \sum_{y \in \mathcal{Y}} P(j|R) Q(y|x_j) \log \frac{Q(y|x_j)}{Q_R(y)}$$

so that $J(Q; \mathcal{R}) = \min_{\{Q_R(\cdot), R \in \mathcal{R}\}} J(Q, Q_R; \mathcal{R})$ recovers the original regularizer.

The local propagation algorithm follows from optimizing each $Q(y|x_i)$ based on fixed $\{Q_R(y)\}$ and subsequently finding each $Q_R(y)$ given fixed $\{Q(y|x_i)\}$. We omit the straightforward derivation and provide only the resulting algorithm: for all points x_i , $i = (l+1) \dots n$ (not labeled), and for all regions $R \in \mathcal{R}$ we perform the following complementary averaging updates

$$Q(y|x_i) \leftarrow \frac{1}{Z_{x_i}} \exp\left(\sum_{R: j \in R} P(R|j) \log Q_R(y)\right) \quad (6)$$

$$Q_R(y) \leftarrow \sum_{j \in R} P(x_j|R) Q(y|x_j) \quad (7)$$

where Z_{x_j} is a normalization constant, and $P(R|j) \propto P(j|R)\gamma(R)$. In other words, $Q(y|x_i)$ is obtained by taking a weighted geometric average of the distributions associated with the regions, whereas $Q_R(y)$ is (as before) a weighted arithmetic average of the conditionals within each region.

Updating $Q(y|x_i)$ for each labeled point x_i , $i = 1 \dots l$ involves minimizing

$$\frac{1}{l} \log Q(y_i|x_i) - \frac{\lambda}{n} H(Q(\cdot|x_i)) - \lambda \sum_{y \in \mathcal{Y}} Q(y|x_i) \left(\sum_{R: j \in R} \gamma(R) P(j|R) \log Q_R(y) \right)$$

where $H(Q(\cdot|x_i))$ is the Shannon entropy of the conditional. While the objective is strictly convex, the solution cannot be written in closed form and has to be found iteratively (e.g., via Newton-Raphson or simple bracketing when the labels are binary). A much simpler update $Q(y|x_i) = \delta(y, y_i)$, where y_i is the observed label for x_i , may suffice in practice. This update results from taking the limit of small λ and approximates the iterative solution.

Thus the transduction information regularization algorithm in the non-parametric setting consists of the following steps:

1. Associate with each region $R \in \mathcal{R}$ a label probability distribution $Q_R(y)$.
2. Initialize $\{Q(y|x_i)\}_{i=1 \dots n}$ and $\{Q_R(y)\}_{R \in \mathcal{R}}$. The initialization values are irrelevant because the objective is convex and admits a unique minimum.
3. Iterate (6) and (7) alternatively until convergence. For labeled points a slightly different update than (6) must be used to account for the observation.

3.1.2 Learning theoretical properties

As in the metric case, we seek to show that the information regularizer is an adequate measure of complexity, in the sense that learning a labeling consistent with a cap on the regularizer requires fewer labeled samples. We consider only the simpler setting where the labels are hard and binary, $Q(y|x_i) \in \{0, 1\}$, and show that bounding the information regularizer significantly reduces the number of possible labelings. Assuming that the points in a region have uniform weights $P(j|R)$, let $N(\gamma)$ be the number of labelings of $\{x_1, x_2, \dots, x_n\}$ consistent with

$$J(Q, \mathcal{R}) < \gamma$$

According to Corduneanu and Jaakkola [2004] we have the following result:

Theorem 4 $\log_2 N(\gamma) \leq C(\gamma) + \gamma \cdot n \cdot t(\mathcal{R}) / \min_R \gamma(R)$, where $C(\gamma) \rightarrow 1$ as $\gamma \rightarrow 0$, and $t(\mathcal{R})$ is a property of \mathcal{R} that does not depend on the cardinality of \mathcal{R} .

Therefore when γ is small, $N(\gamma)$ is exponentially smaller than 2^n , and

$$\lim_{\gamma \rightarrow 0} N(\gamma) = 2$$

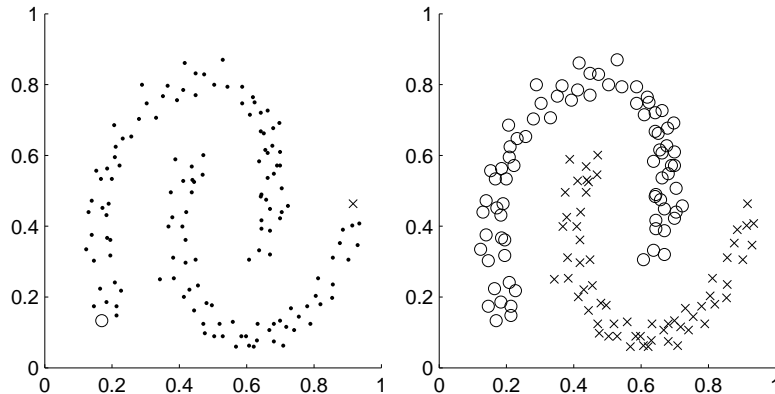


Figure 6: Clusters correctly separated by information regularization given one label from each class

3.1.3 Experiments

To begin with we illustrate the performance of transductive information regularization on two two-dimensional generated binary classification tasks (Corduneanu and Jaakkola [2004]). In this setting we convert the tasks to relational classification by deriving regions of observed points contained in spheres centered at each data point and of a certain radius.

On the classic semi-supervised data set in Figure 6 the method correctly propagates the labels to the clusters starting from a single labeled point in each class. In the example in Figure 7 we demonstrate that information regularization can be used as a post-processing to supervised classification and improve error rates by taking advantage of the topology of the space. All points are a priori labeled by a linear classifier that is non-optimal and places a decision boundary through the negative and positive clusters. Information regularization is able to correct the mislabeling of the clusters. Both results are quite robust to the choice of the radius of the regions as long as all regions remain connected with each other.

Next we test the algorithm on a web document classification task, the WebKB data set of [Blum and Mitchell, 1998]. The data consists of 1051 pages collected from the websites of four universities. This particular subset of WebKB is a binary classification task into 'course' and 'non-course' pages. 22% of the documents are positive ('course'). The dataset is interesting because apart from the documents contents we have information about the link structure of the documents. The two sources of information can illustrate the capability of information regularization of combining heterogeneous unlabeled representations.

Both 'text' and 'link' features used here are a bag-of-words representation of documents. To obtain 'link' features we collect text that appears under all links that link to that page from other pages, and produce its bag-of-words representation.

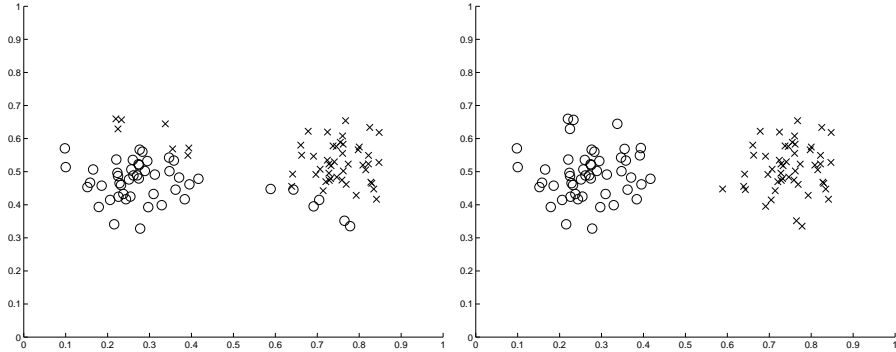


Figure 7: Ability of information regularization to correct the output of a prior classifier (left: before, right: after)

We employ no stemming, or stop-word processing, but restrict the vocabulary to 2000 text words and 500 link words. The experimental setup consists of 100 random selections of 3 positive labeled, 9 negative labeled, and the rest unlabeled. The test set includes all unlabeled documents. We report a naïve Bayes baseline based on the model that features of different words are independent given the document class. The naïve Bayes algorithm can be run on text features, link features, or combine the two feature sets by assuming independence. We also quote the performance of the semi-supervised method obtained by combining naïve Bayes with the EM algorithm as in Chapter ref.chap:Nigam.

We measure the performance of the algorithms by the F-score equal to $2pr/(p+r)$, where p and r are the precision and recall. A high F-score indicates that the precision and recall are high and also close to each other. To compare algorithms independently of the probability threshold that decides between positive and negative samples, the results reported are the best F-scores for all possible settings of the threshold.

The key issue in applying information regularization is the selection of sound relational biases (i.e. \mathcal{R}). For document classification we obtained the best results by grouping all documents that share a certain word into the same region; thus each region is in fact a word, and there are as many regions as the size of the vocabulary. Regions are weighted equally, as well as the words belonging to the same region. The choice of λ is also task dependent. Here cross-validation selected a optimal value $\lambda = 90$. When running information regularization with both text and link features we combined the coverings with a weight of 0.5.

All results are reported in Table 1. We observe that information regularization performs better than naïve Bayes on all types of features, that combining text and link features improves performance of the regularization method, and that on link features the method performs better than the semi-supervised naïve Bayes+EM.

Table 1: Web page classification comparison between naïve Bayes and information regularization and semi-supervised naïve Bayes+EM on text, link, and joint features

	naïve Bayes	inforeg	naïve Bayes+EM
text	82.85	85.10	93.69
link	65.64	82.85	67.18
both	83.33	86.15	91.01

3.2 Parametric classification

We briefly discuss extensions to the transductive information regularization algorithm with relational biases when the conditional takes a parametric form (unpublished work). The extended framework subsumes standard estimation principles such as supervised maximum likelihood, expectation maximization from incomplete data, as well as information regularization presented above. One of the key modifications is to associate with each region R a parametric model $Q_R(x, y|\theta_R)$ instead of the standard average label $Q_R(y)$ as introduced in the above transductive algorithm. With this change the meaning of the regions shifts to represent groups of data points that are modeled in a similar way (same parametric family), where the parametric family may change from region to region. This revision increases the expressive power of information regularization significantly while remaining tractable. Preliminary results are encouraging.

4 Discussion

We have presented the broader information regularization framework, a principle for assigning labels to unlabeled data in a semi-supervised setting. The principle seeks to minimize the information induced between examples and labels relative to a topology over the examples. In other words, we minimize spurious information content not forced by the observed labels.

The information regularization principle manifests itself in different forms depending on assumptions about the space of examples – metric or relational. We demonstrated the resulting algorithms both under the idealized setting where the marginal is known as well as when only a finite unlabeled sample is available. Transductive non-parametric classification results in an efficient algorithm that is provably convergent to a unique optimum.

We can also constrain the conditional probabilities to take a particular parametric form. This extension can be generalized considerably, leading to a unifying framework.

References

- N. Abe, J. Takeuchi, and M. Warmuth. Polynomial learnability of stochastic rules with respect to the KL-divergence and quadratic distance. *IEICE Trans. Inf. & Syst.*, E84-D(3):299–316, March 2001.
- A. Blum, J. Lafferty, M. Rwebangira, and R. Reddy. Semi-supervised learning using randomized mincuts. In *ICML*, 2004.
- R. E. Blahut. Computation of channel capacity and rate distortion functions. In *IEEE Trans. Inform. Theory*, volume 18, pages 460–473, July 1972.
- A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *ICML*. Morgan Kaufmann, 2001.
- A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*. Morgan Kaufmann Publishers, 1998.
- A. Corduneanu and T. Jaakkola. Distributed information regularization on graphs. In *Advances in Neural Information Processing Systems 17*, 2004.
- A. Corduneanu and T. Jaakkola. Continuation methods for mixing heterogeneous sources. In *Proceedings of the Eighteenth Annual Conference on Uncertainty in Artificial Intelligence*, 2002.
- A. Corduneanu and T. Jaakkola. On information regularization. In *Proceedings of the 19th conference on Uncertainty in Artificial Intelligence (UAI)*, 2003.
- Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.
- M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. In S. J. Hanson, G. A. Drastal, and R. L. Rivest, editors, *Computational Learning Theory and Natural Learning Systems, Volume I: Constraints and Prospect*, volume 1. MIT Press, Bradford, 1994.
- M. Seeger. Learning with labeled and unlabeled data. Technical report, Edinburgh University, 2001.
- M. Szummer and T. Jaakkola. Clustering and efficient use of unlabeled examples. In *Advances in Neural Information processing systems 14*, 2001.
- M. Szummer and T. Jaakkola. Information regularization with partially labeled data. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002.

B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Proceedings of the Eighteenth Annual Conference on Uncertainty in Artificial Intelligence*, 2002.