Alin Tomescu

Lecture 14

Linear classifiers and margin

$S_n = \{x_1, x_2, \dots, x_n\}$, $\mathcal{H} =$ set of linear classifiers characterized by θ, θ_0

$$h(x; \theta, \theta_0) = \text{sign}(\theta x + \theta_0)$$

$$h_1 \in \mathcal{H} \text{ predicts} + \ - \cdots +$$

$$h_2 \in \mathcal{H} \text{ predicts} - \ - \cdots +$$

The problem here is that there is no notion of margin incorporated in these.

We'd like to incorporate margin and **say that the only valid labeling is the one with a certain margin**. The larger the margin, the simpler the set of classifiers becomes, because I will have fewer classifier that would be able to satisfy the margin constraints, and so the set becomes smaller.

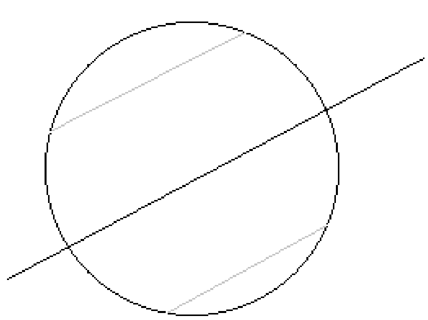
The basic **intuition** is that if we can classify n training examples with a large margin γ, then the classification task is somehow simple.

Definition: When we label with margin γ, we say that $y_1 \dots y_n$ is a valid labeling only if $y_i \frac{\theta x_i + \theta_0}{\|\theta\|} \geq \gamma, \forall i \geq 1$

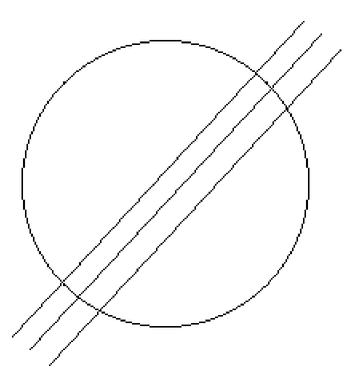
Important note: The notion of margin only makes sense if we specify the margin and *restrict how large the examples can be*. It is the ratio between the two that matters.

Otherwise, consider if someone tells you that they can separate a training set with margin γ, you will not know what that means (see circle drawings). Is it a good or poor result? You need to know how big the circle that encompasses the examples is, in order to tell if a good margin was achieved.

R = 2, \gamma = 1.75



(not drawn to scale) R = 100, \gamma = 1.75

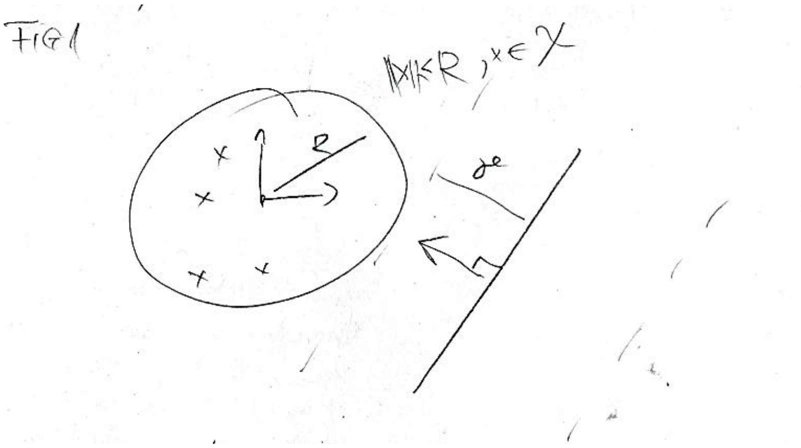


Thus, when dealing with margin, we have to know what R is:

$$R = \max_i \|x^{(i)}\|$$

This way, we only consider training examples: $\|x\| \leq R, x \in \mathcal{X}$.

Question: What is the **minimum** number of possible (not necessarily correct) labelings that the set of linear classifiers can generate over the set of training examples when that has to be done with margin γ?



Answer: $\mathcal{N}_{\mathcal{H}}(S_n; \gamma)$ is always at least one, or even better, always at least two: all points will be + or -. Why isn't the number 2^n? Can't we pick anything? Oh, because a line will either classify everything as + or -

$$\mathcal{N}_{\mathcal{H}}(n; \gamma) = \max_{\substack{x_1 \dots x_n \\ \|x_i\| \leq R}} \mathcal{N}_{\mathcal{H}}(S_n; \gamma)$$

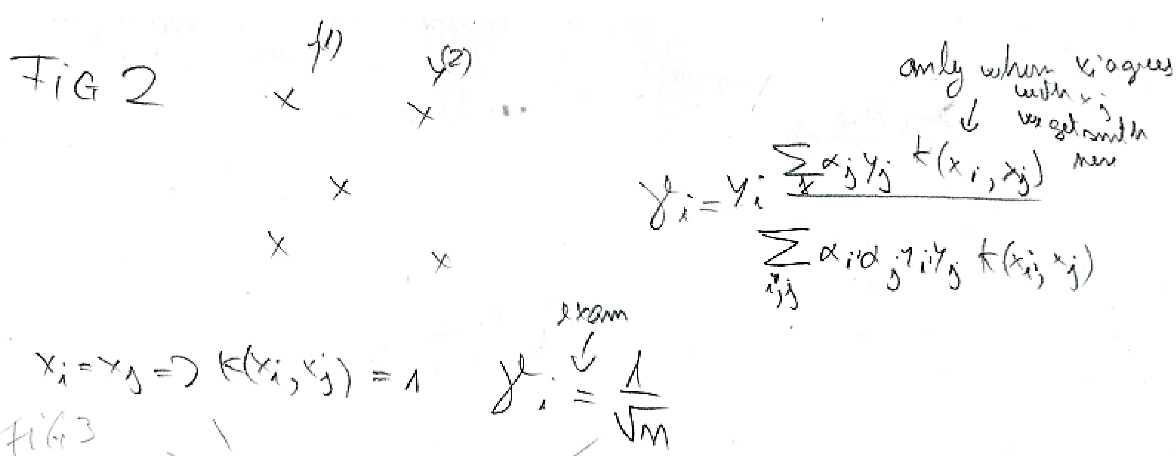
$$d_{VC}(\gamma) = \max\{n : \mathcal{N}_{\mathcal{H}}(n; \gamma) = 2^n\} \leq \min\left\{\frac{R^2}{\gamma^2}, d\right\} + 1$$

Radial basis kernel margin

Let's look at a radial basis kernel, with $\beta > 0$ large.

$$K(x, x') = e^{-\beta \frac{\|x - x'\|^2}{2}}$$

What is the margin that I can attain over an arbitrary labeled set of points as $\beta \to \infty$? (What is γ?)

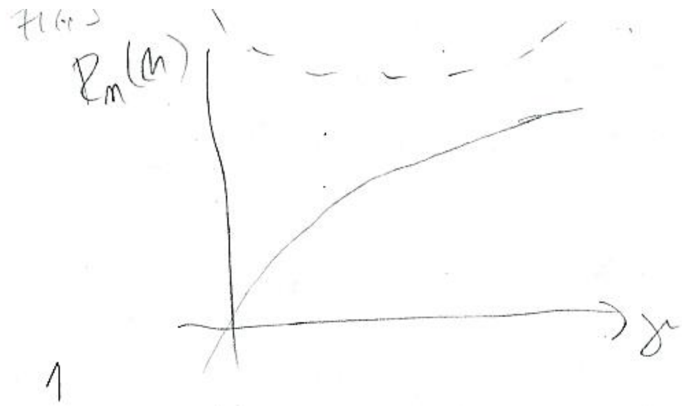


Explanation: If β is large, then only when x_i “agrees”with x_j, we get $K(x_i, x_j)$ equal to 1.

$$\gamma_i = y_i \frac{h(x_i;\theta)}{\|\theta\|} = y_i \frac{\sum_j \alpha_j y_j K(x_i, x_j)}{\sqrt{\sum_k \sum_j \alpha_k \alpha_j y_k y_j K(x_k, x_j)}} = (\text{large } \beta) = \frac{\alpha_i}{\sqrt{\sum_j \alpha_j^2 y_j^2 K(x_j, x_j)}} = \frac{\alpha_i}{\sqrt{\sum_j \alpha_j^2}}$$

$$= (YKY \text{ is } I_n, \text{see HW2, prob. 1, part e}) = \frac{1}{\sqrt{n}}$$

What is R in this case? $R = 1$ because all feature vectors have norm 1 when using an RBF (they appear in the surface of this infinite dimensional hypersphere). $R^2 = \|\phi(x)\|^2 = K(x, x) = 1$.



Generalization bounds that depend on margin

In the previous lecture, we showed that for all classifiers the following holds with probability at least $1 - \delta$:

$$\forall h \in \mathcal{H}, R(h) = R_n(h) + \sqrt{\frac{d_{VC}\left(1 + \log\left(\frac{2n}{d_{VC}}\right)\right) + \log\left(\frac{4}{\delta}\right)}{n}}$$

Note: Not really important to understand how this came to be, but more important to understand how margin and generalization interplay.

large margin ↑ ⇒ smaller generalization error ↓

How do we change this such that we can incorporate the margin?

$$d_{VC} \le \min\left\{\frac{R^2}{\gamma^2}, d\right\} + 1$$

$$R_n(h;\gamma) = R_n(\theta, \theta_0; \gamma) = \sum_i [[\frac{y_i(\theta x_i + \theta_0)}{\|\gamma\|} < \gamma]]$$

(This was not covered in class but was expanded upon in HW6, as can be seen below)

For linear classifiers in a feature space, where $\|\phi(x)\| \leq 1$ we have $d_{VC} \leq \frac{1}{\gamma^2}$, we can replace d_{VC} by its upper bound and obtain the following:

$$\forall \theta, R(\theta) \leq R_n(\theta;\gamma) + \sqrt{\frac{\frac{1+\log(2n\gamma^2)}{\gamma^2} + \log\left(\frac{4}{\delta}\right)}{n}}$$

Note: We are skipping the part about generalizations on distributions of classifiers.

Reconstructing the underlying distribution of the training data

So far, we've talked about discriminative methods. We never explicitly reconstructed the underlying distribution of the training examples.

- Supervised learning case (simple)
  - Reconstruct $p^*(x,y) = p^*(x|y)p^*(y)$ from $(x_i, y_i), i = 1, \ldots, n$
- Unsupervised learning case (harder)
  - Reconstruct $p^*(x,y) = p^*(x|y)p^*(y)$ from $x_i \sim p_x^*, i = 1, \ldots, n$
- Semi-supervised learning case
  - Reconstruct $p^*(x,y) = p^*(x|y)p^*(y)$ from $x_i \sim p_x^*$ and (x_j, y_j)

Supervised learning case (simple)

We are given $S_n = \{(x_i, y_i), 1 \leq i \leq n\}$ and we want to put two Gaussians on points: one on the plus points and one on the minus points.

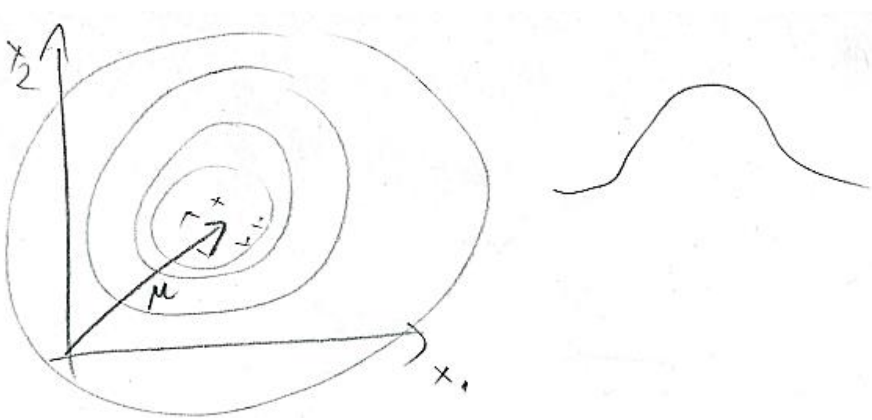
What is the first step? What is the first task we have to define? We need to assume some underlying set of possible distributions.

1) Parameterize $P(x, y; \theta), \theta \in \Theta$
2) Estimate these probabilities (ML, MAP, Bayesian)

Let's look at how we can do this with Gaussian distributions.

Reminder: Gaussian looks like this:

FIG 4

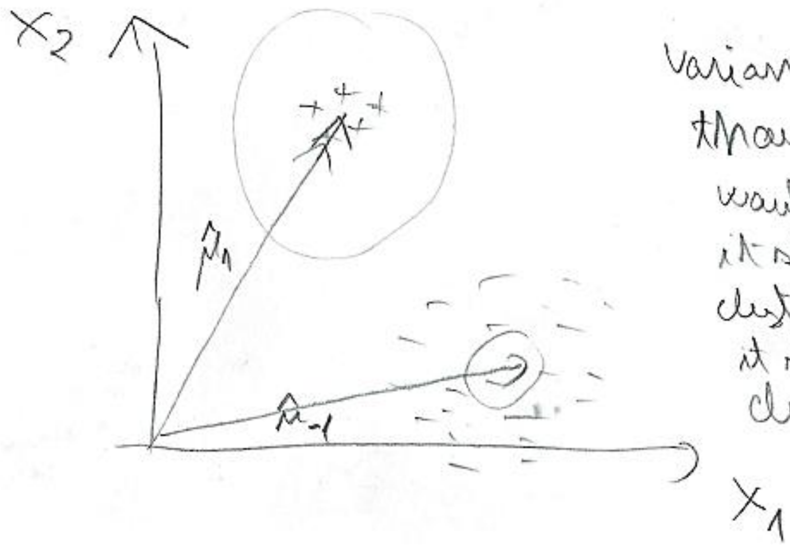


$$p(x|y;\theta) = \mathcal{N}\left(x;\mu_y,\sigma^2 I\right), \sigma \text{ is same for each } y = \pm 1$$

$$p(y;\theta) = p_y, p_1 + p_{-1} = 1$$

The parameters will be:

$$\theta = [\mu_1, \mu_{-1}, \sigma^2, p_1, p_{-1}]$$



Variance would be shared though, so variance would be larger than it should be on the + cluster and smaller than it should be on the − cluster

Note: Variance is shared between $p(x|1;\theta)$ and $p(x|-1;\theta)$, so for the particular example above, variance would be larger than it should be on the + **cluster** and smaller than it should be on the − **cluster**.

How to estimate Gaussian?

$$D = \left\{x^{(i)}, i = 1, \dots, n\right\} \text{ maximum likelihood (ML) estimation}$$

We look at the log-likelihood of $P(x|y)$:

$$l(\mu,\sigma^2;D) = \sum_{i=1}^{n} \log \mathcal{N}\left(x^{(i)};\mu,\sigma^2 I\right) = \sum_{i=1}^{n} \left[-\frac{1}{2\pi\sigma^2}\left\|x^i - \mu\right\|^2 + \frac{d}{2}\log(2\pi\sigma^2)\right]$$

Note: Remember we are dealing with a multivariate Gaussian, so there's a covariance matrix ($\sigma^2 I$) determinant somewhere in there that gives us the $\frac{d}{2}\log(2\pi\sigma^2)$

$$\frac{\partial}{\partial\mu} l(\mu,\sigma^2;D) \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

$$\frac{\partial}{\partial\sigma^2} l(\mu,\sigma^2;D) \Rightarrow \widehat{\sigma^2} = \frac{1}{nd}\sum_{i=1}^{n} \left\|x^{(i)} - \hat{\mu}\right\|^2$$

$$\widehat{\sigma^2} = \frac{1}{nd}\sum_{i=1}^{n}\left\|x^{(i)} - \hat{\mu}\right\|^2 = \widehat{\sigma^2} = \frac{1}{nd}\left(\sum_{i:y^{(i)}=1}^{n}\left\|x^{(i)} - \widehat{\mu_1}\right\|^2 + \sum_{i:y^{(i)}=-1}^{n}\left\|x^{(i)} - \widehat{\mu_{-1}}\right\|^2\right)$$