

Lecture 3: Linear classifiers continued

Maximum margin linear classifier: $\min_{\theta} \frac{1}{2} \|\theta\|^2$, such that $y^{(i)}\theta x^{(i)} \geq 1, \forall i = 1, \dots, n$

Properties of this solution:

- $\hat{\theta}$ is unique
- Solution is sparse: we rely only on the subset of the training example
 - $y^{(1)}(\hat{\theta}x^{(1)}) = 1$, these examples are called support vectors
 - Points that lie exactly on the boundary margin
 - $y^{(2)}(\hat{\theta}x^{(2)}) > 1$
 - ...
 - Sparsity leads to compression
- Solution is very sensitive to outliers. The boundary can shift dramatically with an outlier.
 - If we have mislabeled training examples, for instance

Support vector machines

We have a maximum margin linear classifier (MMLC), that we call a support vector machine (SVM)

$$\min_{\theta, \theta_0} \frac{1}{2} \|\theta\|^2, \text{ such that } y^{(i)}(\theta x^{(i)} + \theta_0) \geq 1, \forall i = 1, \dots, n$$

How do we accommodate violations of MMLC constraints? Like allowing examples that might be correctly classified but not on the right of the margin boundary...

$$\min_{\theta, \theta_0, \xi_i} \frac{1}{2} \|\theta\|^2 + \frac{C}{n} \sum_{i=1}^n \xi_i, \text{ such that } y^{(i)}(\theta x^{(i)} + \theta_0) \geq 1 - \xi_i, \forall i = 1, \dots, n, \xi_i \geq 0$$

ξ_i is called a **slack variable**:

- $\xi_i > 0$ for points correctly classified after the boundary
- $\xi_i \in (0,1)$ for points almost correctly classified between the boundary and the line
- $\xi_i > 1$ for points incorrectly classified

As C increases ($C \rightarrow \infty$), we prefer **less slack** at the cost/expense of a **smaller margin**.

As C decreases, we prefer a **larger margin** at the cost of **more slack**.

$C = 0$ means there is **no cost to mislabeling** negative points so they would not appear as support vectors (i.e. points that the solution would depend on). You can give the points as much slack as you want and you can still satisfy the constraints without affecting the value of θ or θ_0 .

This is still a quadratic programming problem.

If we fix θ, θ_0 . What is the **maximum slack**? The minimum would be:

$$\hat{\xi}_i(\theta, \theta_0) = 1 - y^{(i)}(\theta x^{(i)} + \theta_0)$$

Alin Tomescu

6.867 Machine learning | Week 2, Thursday, September 10th, 2013 | Lecture 3

Assuming $y^{(i)}(\theta x^{(i)} + \theta_0)$ is positive, since if:

$$y^{(i)}(\theta x^{(i)} + \theta_0) \in (0,1)$$

Then, we need to “allow” this i^{th} example to be inside the maximum margin. It needs $1 - y^{(i)}(\theta x^{(i)} + \theta_0)$ slack.

$$\hat{\xi}_i(\theta, \theta_0) = \max\{1 - y^{(i)}(\theta x^{(i)} + \theta_0), 0\}$$

So,

$$\min_{\theta, \theta_0, \xi_i} \frac{1}{2} \|\theta\|^2 + \frac{C}{n} \sum_{i=1}^n \hat{\xi}_{i_i}(\theta, \theta_0)$$

$\hat{\xi}_{i_i}(\theta, \theta_0)$ looks like a loss on the i^{th} training example

$\frac{1}{2} \|\theta\|^2$ is the regularization

Balance between regularization and loss on the training examples is a recurring theme in machine learning.

$$z_i = 1 - y^{(i)}(\theta x^{(i)} + \theta_0) = \text{agreement}$$

$$\text{Loss}_h(z_i)$$

Non-linear predictors

$$X \rightarrow Y$$

So far we've looked at $Y = \{-1, 1\}$ and $X = \mathbb{R}^d$

Y could be $\{1, \dots, k\}, \mathbb{R}, \{-1, 1\}^m, G \in (V, E), \{\emptyset, 1\}$ a.k.a. 1-class

X could be a document, a trajectory, a graph.

Consider $X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbb{R}^2$

We can remap the input: $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \rightarrow \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_1 x_2 \sqrt{2} \\ x_1 \\ x_2 \end{bmatrix} = \phi(x) \in \mathbb{R}^5$

Take each training example and map it to a feature vector: $x^{(i)} \rightarrow \phi(x^{(i)}) = y^{(i)}$

$$h(x; \theta, \theta_0) = \text{sign}(\theta \phi(x) + \theta_0), \theta \in \mathbb{R}^5$$

$$\theta \phi(x) + \theta_0 = \theta_1 x_1^2 + \theta_2 x_2^2 + \theta_3 x_1 x_2 \sqrt{2} + \theta_4 x_1 + \theta_5 x_2 + \theta_0$$

$$\phi(x) \phi(x') = (xx') + (xx')$$