Support Vector Machines

Michael Collins

January 22, 2012

The Maximum-Margin Classifier

- For a given training set $(\underline{x}_t, y_t)$ for $t = 1 \ldots n$
- For a given parameter vector $\underline{\theta}$:
  - The *functional margin* on the t'th example is

$$\gamma_t(\underline{\theta}) = y_t(\underline{\theta} \cdot \underline{x}_t)$$

    NOTE: if $\gamma_t(\underline{\theta}) > 0$ for all t, then $\underline{\theta}$ correctly classifies all training examples
  - The *geometric margin* on the t'th example is

$$\gamma_t(\underline{\theta})/||\underline{\theta}||$$

    This is the distance of the t'th point to the hyperplane (a negative distance means the point is classified incorrectly)
  - The *geometric margin* on the training set is then

$$\gamma(\underline{\theta}) = \min_t \gamma_t(\underline{\theta})/||\underline{\theta}||$$

The Maximum-Margin Hyperplane

- Assume that the data is *separable*, i.e., there exists a hyperplane that correctly classifies all training points. The maximum-margin hyperplane is defined by $\underline{\theta}^*$, where

$$\underline{\theta}^* = \arg\max_{\underline{\theta}} \gamma(\underline{\theta})$$

  It has a geometric margin on the training set of $\gamma(\underline{\theta}^*) = \gamma^*$

- The perceptron convergence result: assume in addition that for all t, $||\underline{x}_t||^2 \leq R^2$ for some constant R. Then the perceptron algorithm makes at most

$$\frac{R^2}{\gamma^{*2}}$$

  mistakes before convergence

Finding the Maximum-Margin Classifier (the *Support Vector Machine*)

- An optimization problem:
  Find the value for $\underline{\theta}$ that minimizes

$$\frac{1}{2}||\underline{\theta}||^2$$

  subject to the constraints

$$y_t(\underline{x}_t \cdot \underline{\theta}) \geq 1 \quad \text{for all } t = 1 \ldots n$$

- The solution is the maximum margin classifier $\underline{\theta}^*$
- This is a *quadratic programming problem*: optimization of a quadratic objective with linear constraints

Adding a Bias (Offset) Parameter

- An optimization problem:
  Find the value for $(\underline{\theta}, \theta_0)$ that minimizes

$$\frac{1}{2}||\underline{\theta}||^2$$

  subject to

$$y_t(\underline{x}_t \cdot \underline{\theta} + \theta_0) \geq 1 \quad \text{for all } t = 1 \ldots n$$

- Note: the bias parameter θ_0 only appears in the constraints

Benefits of the Maximum Margin Solution

- The maximum margin solution for a given training set is unique
- Intuition: drawing the separating hyperplane as far as possible from the training examples will lead to good generalization properties (we'll see some formal guarantees later)

Support Vectors

- The maximum margin hyperplane only depends on a subset of the training examples, namely those examples that appear exactly on the margin. These points are called *support vectors*

A Problem: Sensitivity to Outliers

- If the training data is not separable, the maximum-margin hyperplane does not exist (the optimization problem has no solution)
- Even a single training example can radically change the position of the maximum margin classifier

Introducing Slack Variables

minimize (with respect to $\underline{\theta}, \theta_0$, and ξ_t for $t = 1 \ldots n$)

$$\frac{1}{2}||\underline{\theta}||^2 + C\sum_{t=1}^{n}\xi_t$$

subject to

$$y_t(\underline{\theta} \cdot \underline{x}_t + \theta_0) \geq 1 - \xi_t \text{ and } \xi_t \geq 0 \text{ for all } t = 1 \ldots n$$

- ξ_t is a "slack variable" for the t'th example
- $C > 0$ is a constant