Support Vector Machines

The Parzen density is a linear machine in the kernel feature space; which Parzen kernels satisfy the Mercer condition? Clearly, not all linear machines in feature space are Parzen densities, for example the linear kernel does not produce a valid density estimate. All Parzen kernels need to satisfy the integration to one and positivity condition; are these sufficient for the Mercer condition? The weights for the Parzen density must be non-negative and sum to one, so a natural generalization is to allow the weights to be any convex combination. Since the Parzen density is a linear machine in an RKHS, what regularized cost function does it minimize? The Parzen classifier is the mean of the points in feature space, which minimizes the sum of squared distances to the points. The Parzen classifier maximizes the functional margin with weight $1/n$ for each point, and weight $1/2$ for the weight norm penalty.

SVM has the constraint

$$\sum_i y_i \alpha_i = 0, \tag{1}$$

which makes the total weight for the positive class equal to that of the negative class. The Parzen classifier automatically satisfies this constraint since the total weight for each class is one.

1 Separable Case

We are given data $(\mathrm{x}_i, y_i), \quad x \in R^d, y \in \{-1, 1\}$. We want a linear classifier in an infinite-dimensional kernel space,

$$g(x) = sign(\phi(w) \cdot \phi(x) + b), \tag{2}$$

where

$$\phi(w) \cdot \phi(x) = K(w, x). \tag{3}$$

The SVM optimization is $\phi^*(w) = \arg\min_{\phi(w)} \frac{1}{2} \phi(w)^2$, $suchthat \quad y_i(\phi(w) \cdot \phi(x_i) + b) \geq 1$. So of all the classifiers which correctly classify the data, we want the one closest to the origin. From a Bayesian perspective, this is choosing the most probable classifier under a zero mean normal prior, which has likelihood above a certain threshold.

The Lagrangian is

$$L(\phi(w), b, \alpha) = \frac{1}{2} ||\phi(w)||^2 - \sum_i \alpha_i (y_i(\phi(w) \cdot \phi(x_i) + b) - 1). \tag{4}$$

The stationary conditions are $\mathrm{L}(\phi(w), b, \alpha)\phi(w) = \phi(w) - \sum_i y_i\alpha_i\phi(x_i) = 0$, $L(\phi(w), b, \alpha)b = \sum_i y_i\alpha_i = 0$. So the weight vector is a linear combination of the data points:

$$\phi(w) = \sum_i y_i\alpha_i\phi(x_i). \tag{5}$$

The classifier is then $\mathrm{g(x)} = \mathrm{sign}\Bigg(\sum_i y_i\alpha_i\phi(x_i)\cdot\phi(x) + b\Bigg)$ $= sign\Bigg(\sum_i y_i\alpha_i K(x_i, x) + b\Bigg)$. Substituting back into the Lagrangian gives the dual cost function $\mathrm{W}(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} y_iy_j\alpha_i\alpha_j\phi(x_i)\cdot\phi(x_j)$ $= \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} y_iy_j\alpha_i\alpha_jK(x_i, x_j)$. The optmization is now $\hat{\alpha} = \arg\max_\alpha W(\alpha)$, $such that \quad \alpha_i \geq 0$.

1.1 Kernel Density Connection

We can write the SVM classifier as $\mathrm{g(x)} = \mathrm{sign}\Bigg(\sum_i y_i\alpha_iK(x_i, x) + b\Bigg)$ $= sign\Bigg(\sum_{i:y_i=1}\alpha_iK(x_i, x) - \sum_{j:y_j=-1}\alpha_jK(x_j, x) + b\Bigg)$ $= sign\Bigg(h_+(x) - h_-(x) + b\Bigg)$.

2 Parzen to Kernel SVM

Let us now start from a Parzen density perspective. For a two class problem we can use the following discrimant:

$$s(x) = sign[p(x|1) - p(x|-1)], \tag{6}$$

by assuming equal class priors $p(1) = p(-1)$. We estimate the class conditional densities use Parzen estimates so that:

$$p(x|1) - p(x|-1) = \frac{\sum_i \beta_iy_iK(x, x_i)}{2\sum_i \beta_i}, \tag{7}$$

where $\beta_i \geq 0$,
$\sum_i \beta_iy_i = 0$, Essentially we are picking weights or a distribution of the examples while remaining consistent with the equal class priors assumption.

Now the margin of an example under this discriminant is

$$m_i = y_is(x_i) = y_i[p(x_i|1) - p(x_i|-1)], \tag{8}$$

which is a measure of "how correctly" the example is classified. In other words, large and positive margins correspond to confident and correct classifications.

Now we can write the expected margin of the examples under the Parzen density estimates as:

$$E[m] = \frac{\sum_i \beta_i y_i \sum_j \beta_j y_j K(x_i, x_j)}{2 \sum_{i=1}^n \beta_i} = \frac{\sum_{i,j} \beta_i \beta_j y_i y_j K(x_i, x_j)}{2 \sum_{i=1}^n \beta_i}. \tag{9}$$

The SVM solution seems to choose α such that the worse distribution in terms of classification results, namely a distribution centered on examples lying on the margin. Recall that in feature space, the SVM chooses the maximal margin hyperplane that statisfies the classification constraints. This corresponds to minimizing the norm of the weight vector. In the kernel SVM, the norm of the weight vector is the expected margin. Thus in a structural risk minimization fashion, the SVM is choosing the weakest Parzen classifier that still correctly classifies the training data.