10-701/15-781, Machine Learning: Homework 5

Eric Xing, Tom Mitchell, Aarti Singh
Carnegie Mellon University
Updated on March 24, 2010

- The assignment is due at 10:30am (beginning of class) on **Mon, April 26, 2010**.
- Separate you answers into three parts, one for each TA, and put them into 3 piles at the table in front of the class. Don't forget to put both your name and a TA's name on each part.

1 SVMs [Amr, 30 + 10 extra credits points]

1.1 Kernels and Feature Maps

Given the following dataset in 1-d space (Figure 2), which consists of 3 positive data points $\{-1, 0, 1\}$ and 3 negative data points $\{-3, -2, 2\}$.

Figure 1: Dataset

(1) **(3 pts)** Find a feature map($\{\mathbf{R}^1 \rightarrow \mathbf{R}^2\}$), which will map the data in the ordinal 1-d *input space* (x) to a 2-d *feature space* (y_1, y_2) so that the data becomes linearly-separable. Plot the dataset after mapping in 2-d space.

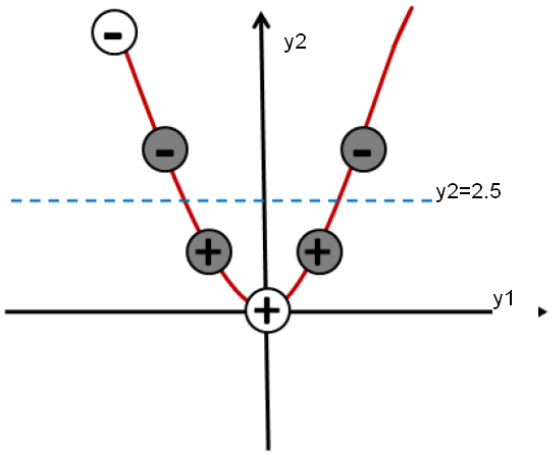

Figure 2: Feature Space

The feature map is $y_1 = x$ and $y_2 = x^2$.

(2) **(5 pts)** Write down the decision boundary $w_2y_2+w_1y_1+w_0$ given by hard-margin linear SVM in the feature space. Draw this decision boundary on your plot and mark the corresponding support vector(s).

$y_2 = 2.5$

(3) **(5 pts)** What is the equivalent decision boundary in the original input space? Draw this decision boundary over the points in Figure 2.

$x^2 - 2.5 = 0$, which is $x = \pm\sqrt{2.5}$ and decision boundary is the interval $[-\sqrt{2.5}, \sqrt{2.5}]$: inside this interval the prediction is positive and outside it the prediction is negative. Note that the input space is 1-d thus the decision boundary is an interval.

(4) **(3 pts)** For the feature map you choose, what is the corresponding kernel $K(x_1, x_2)$?

$K(x_1, x_2) = x_1x_2 + x_1^2x_2^2$.

(5) **(4 pts)** What is the maximum number of points in the input space that can be shattered by an SVM classifier using the kernel in (4)? Explain in one or two sentences.

This is just the VC-dimension of the SVM with the kernel in (4). The Kernel maps the data to a 2-d space, and since SVM learns a linear decision boundary in the feature space, the VC-dimension is $2 + 1 = 3$

1.2 SVM and other Classifiers

(1) The idea of mapping the data from the input space to another feature-space in which the data becomes linearly separable was used earlier in the class by another classifier.

(a) **(2 pts)** What is the name of this classifier?

Neural Network. Recall in HW3 when you were asked to draw the input after the hidden layer and you saw that the hidden layer maps the data into a space in which the data becomes linearly separable.

(b) **(3 pts)** List one difference between the way this idea was used by SVM and this classifier. Discuss the implication of this difference on both classifiers (with regard to training and/or testing, etc.).

NN learns the mapping function, and as such the optimization problem is not-convex and all computations depend on the dimensionality of the feature space, however in SVM the user supplies the Kernel function which implies the feature map. However, in SVM the optimization problem is convex and all computations are performed in the input

space (the low-dimensional space).

(2) The final decision rule of an SVM classifer using kernel K is given by:

$$y*(z) = \text{sign}\Big(\sum_{i \in \text{SV}} \alpha_i y_i K(x_i, z) + b\Big) \tag{1}$$

In this case K can be interpreted as a similarity metric.

(a) **(2 pts)** We have studied one classifier whose decision rule looks similar to (1), what is the name of this classifier?

K-Nearest Neighbor.

(b) **(3 pts)** List one difference between the classifier in (a) and SVM with regard to their decision rules and discuss the implication of this difference on both classifiers (again with regard to training and/or testing, etc).

SVM learns a global decision rule and as such the weights (αs) assigned to each of the points are learned via a global optimization problem (the dual), and few points are selected (support vectors). However, in K-NN, the points that affect the prediction of a new test point z are different for every z and are the K-nearest neighbor to z (i.e. the decision rule is local for every point).

1.3 EXTRA CREDIT: Dimension of Transformed Feature Space

One popular choice for the kernel, K, in SVM is the dth degree polynomial:

$$K(\mathbf{x}, \mathbf{z}) = (1 + \langle \mathbf{x}, \mathbf{z} \rangle)^d$$

The 2nd degree polynomial kernel corresponds to the inner product of a transformed feature space. If $\mathbf{x}$ and $\mathbf{z}$ are both 2-dimensional vectors in the input space , then the dimension of the transformed feature space is 6:

$$\begin{aligned} k(\mathbf{x}, \mathbf{z}) &= (1 + \langle \mathbf{x}, \mathbf{z} \rangle)^2 \\ &= (1 + x_1 z_1 + x_2 z_2)^2 \\ &= 1 + 2x_1 z_1 + 2x_2 z_2 + x_1^2 z_1^2 + 2x_1 z_1 x_2 z_2 + x_2^2 z_2^2 \\ &= \langle [1 \ \sqrt{2}x_1 \ \sqrt{2}x_2 \ \ x_1^2 \ \sqrt{2}x_1 x_2 \ \ x_2^2], [1 \ \sqrt{2}z_1 \ \sqrt{2}z_2 \ \ z_1^2 \ \sqrt{2}z_1 z_2 \ \ z_2^2] \rangle \\ &= \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle \end{aligned}$$

(1) **(5 pts)** Let's generalize this example. Assume $\mathbf{x}$ and $\mathbf{z}$ are p-dimensional, what is the dimension of the transformed feature space for the dth degree polynomial kernel? **Note**: you don't have to explicitly list the form of each dimension as in the example above, I am just looking for the dimensionality. (**Hint:** This is a combination with replacement problem — brush up on your combinatorics background)

This question is a combination with repetition problem where you can choose one of $(p+1)$ different objects each time and you have to make this choice d times. The solution is:

$$\left(\begin{array}{c} (p+1)+d-1 \\ d \end{array} \right) = \left(\begin{array}{c} p+d \\ d \end{array} \right)$$

(2) **(5 pts)** According to (1), calculate the dimension of the transformed feature space for a 15-dimensional $\mathbf{x}$ using the 3rd degree polynomial kernel. Also for this specific setting, compute the ratio between the number of multiplications required to evaluate the kernel in the input space vs. evaluating the dot-product in the feature space.

Dimension of feature space = 816.
ratio = $\frac{15}{816} = .018$.