2018 Spring STA 561: Homework 4

Duke University

February 17, 2018

1 Constructing Kernels (15 pts)

In class, we saw that by choosing a kernel

$$k(x_i, x_k) =< \Phi(x), \Phi(x) >_{\mathcal{H}_k},$$

we can implicitly map data to a high dimensional space, and have the SVM algorithm work in that space. One way to generate kernels is to explicitly define the mapping ϕ to a higher dimensional space, and then work out the corresponding K. However in this question we are interested in direct construction of kernels. Let K_1 be kernels over $\mathbb{R}^n \times \mathbb{R}^n$, let $a \in \mathbb{R}$. $< a, b >$ denotes the dot product, $a^T b$.

For each of the function K below, state whether it is necessarily a kernel. If you think it is, prove it; Otherwise, please specify reasons.

(a) $K(x, z) = aK_1(x, z)$
(b) $K(x, z) =< x, z >^3 +(< x, z > -1)^2$
(c) $K(x, z) =< x, z >^2 + \exp(-\|x\|^2) \exp(-\|z\|^2)$

2 Reproducing kernel Hilbert spaces (20 pts)

Let $\mathscr{F}$ be the set of all functions $f : [0, 1] \to \mathbb{R}$ such that $f(x) = ax$ for some real number a. Show that this is a RKHS with kernel $K(x, y) = xy$.

3 Convexity and KKT conditions (40 pts)

In class, we have seen the hinge loss that is used for "maximum-margin" classification, most notably for support vector machines. In this problem, you will see a new loss function that is defined as follows.

$$L(x, y, f) = \max(0, |y - f(x)| - \epsilon).$$

This loss is called the "epsilon-insensitive loss," and we hope you can see why it has this name! This is a very popular loss function. The cost function for the SVMs will thus be:

$$\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} L(x_i, y_i, f)$$

where x is the input, y is the output, and $f(x) = w^T x$ is used for predicting the label. Both C and $\epsilon > 0$ are parameters.

Hint: The primal form for this problem is given as:

$$\min_{w,\eta,\eta*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\eta_i + \eta_i^\star)$$

subject to

$$y_i - \langle w, x_i\rangle - \epsilon \leq \eta_i$$
$$\langle w, x_i\rangle - y_i - \epsilon \leq \eta_i^\star$$
$$\eta_i, \eta_i^\star \geq 0, i = 1, \ldots, n,$$

where $\eta, \eta^\star$ are two slack variables.

(a) Please write down the Lagrangian function for the above primal form, and use KKT conditions to derive the dual form. You can look up the answer on the internet if you get stuck but try it first so you get used to doing this type of calculation.

(b) How would you define support vectors in this problem?

(c) How does ϵ influence the complexity of the model in practice? In other words, does increasing ϵ make the model more or less likely to overfit in general?

(d) How does C influence the complexity of the model in practice? In other words, does increasing C make the model more or less likely to overfit in general?

(e) Now suppose you are given a new (unseen) sample x, please write down the equation for evaluating $f(x)$.

4 SVM Implementation (25 pts)

In this problem, you will experiment with the support vector machine (SVM), and various kernel functions.

(a) Implement a "hard" maximum-margin SVM classifier in MATLAB, R, or Python. Here, "hard" means that if the data are separable the SVM will return a maximum margin solution, but if the data are not separable, the code will fail. Your implementation should optimize the dual problem using a quadratic program solver or a specialized solver, and include a *train* function and a *predict* function. (Note that you do not need to write the solver yourself!)

(b) Train an SVM classifier with the kernel function $k(\mathbf{x}, \mathbf{z}) = \mathbf{x}^\top \mathbf{z}$ on 9/10ths of the credit card data set. (Please set seed as 2018 to choose which tenth to leave for testing.) What is the accuracy of this classifier on the test data set? Show the ROC curves, and also report the AUC.

(c) Train an SVM classifier with the radial basis kernel

$$k(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{z}\|_2^2}{\sigma^2}\right)$$

on the credit card data training set, for $\sigma^2 = 5$ and $\sigma^2 = 25$. What is the accuracy of these classifiers on the test data set? Show the ROC curves, and also report the AUC.