CS771: Machine learning: tools, techniques, applications
Assignment #2: SVM, Kernels, Regression

Due on: 16-3-2016, 23.59 07-3-2016
MM: 200

1. *Hinge loss* is often used as the loss function for maximum margin classification. It is defined as:

   $$L(y) = \max(0, 1 - t \cdot y)$$

   here $t = \pm 1$ the intended output and y is the actual raw output from the decision function (say $\mathbf{w}^T\mathbf{x} + w_0$). Notice that if $|y| \geq 1$ and the label is correct that is t and y have the same sign then $L(y) = 0$ otherwise it is increasing linear in y. Note that hinge loss is a convex function.

   For the spam data set you used earlier use the hinge loss function in the SVM classifier and compare the results you get (5-fold cross validated) with the standard formulation. [30+20=50]

2. You are given a dataset of Connect Four game positions and the final outcome (win/loss/draw) for the first player. In each of the game positions, only 8 moves have been made so far with none of the players having won yet and the next move isn't forced.

   The dataset is at: https://archive.ics.uci.edu/ml/datasets/Connect-4

   Report 5 fold cross validation results. Try the following approaches using an SVM:

   1. One-Versus-Rest
   2. One-Versus-One

   For a list of SVM libraries available in different languages, have a look at:
   http://www.support-vector-machines.org/SVM_soft.html
   `You should not directly use the multiclass classification option of these SVM libraries`.
   (Hint: For using SVM, change the dataset appropriately. E.g., use 42×3 features instead of 42 as present in the dataset. For every i^{th} feature in the dataset, if the feature is o then set $3 \times i$ as 1, if b then $3 \times i - 1$ as 1 and if x then $3 \times i - 2$ as 1 and rest are set to 0. Also for class labels, you can use nominal 1 for win, 0 for draw and -1 for a loss.)

   [60]

3. We saw closure properties allowed new kernels to be created from existing kernels. Prove the statements below regarding these closure properties, or give counter-examples to disprove them. Assume $\mathbf{x}, \mathbf{z} \in \mathcal{X} = \mathbb{R}^d$.

   (a) If K_1 is a kernel on $\mathcal{X}$, then $K(\mathbf{x}, \mathbf{z}) = e^{K_1(\mathbf{x}, \mathbf{z})}$ is also a kernel.

   (b) $K(\mathbf{x}, \mathbf{z}) = e^{(\|\mathbf{x}\|^2 + \|\mathbf{z}\|^2)} \cdot (\frac{\mathbf{x}^T\mathbf{z}}{\|\mathbf{x}\|^2\|\mathbf{z}\|^2})$ is a kernel.

   (c) $K(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{d} min(|\mathbf{x}_i|, |\mathbf{z}_i|)$ is a kernel

   [30]

4. Consider a regression problem, whereby, we are given feature vectors $\{\mathbf{x}_i \in \mathbb{R}^d\}$ and response variables $\{y_i \in \mathbb{R}\}$. The objective is to minimize the error between the estimated and true response variables. In order to control overfitting, we add a regularization term. The problem can be formulated as follows:

$$\begin{aligned} \underset{\mathbf{w},\boldsymbol{\xi}}{\text{minimize}} \quad & L = \sum_{i=1}^{n} \xi_i^2 \\ \text{subject to} \quad & y_i - \mathbf{w}^T\mathbf{x} = \xi_i, \ \forall i = 1, 2, \ldots n \\ & \|w\|_2 \leq B. \end{aligned}$$

Here, B is the regularization parameter.

(a) Obtain a solution of the problem by rewriting it in dual form.

(b) Does this problem have the equivalent of support vectors as in SVMs? Justify.

(c) What is one basic disadvantage of the above as compared to the SVM solution?

[30+25+5=60]