Assignment 4: Kernels and SVM's

Submission: Thursday November 4th
Maximum of 2 students per group

Prof. Fabio A. González
Machine Learning - 2010-II
Maestría en Ing. de Sistemas y Computación

---

1. Regression on strings

   (a) Implement a function that calculates a kernel over fixed-length strings,

   $$k : \Sigma^d \times \Sigma^d \quad \rightarrow \quad \mathbb{R},$$

   which counts the number of **coincidences** between two strings.

   (b) Implement the kernel ridge-regression (KRR) algorithm.

   (c) Use the KRR implementation and the kernel k to train a model using the training data set in `http://dis.unal.edu.co/~fgonza/courses/2008-I/ml/assign4-train.txt`. Evaluate the error of the model on the training data set. Plot the output of the model on the training data along with the real output values (results may be sorted by the real output value).

   (d) Evaluate the trained model on the test data set `http://dis.unal.edu.co/~fgonza/courses/2008-I/ml/assign4-test.txt`. Plot the results and discuss them.

   (e) Build a new kernel, k', composing the kernel k with more complex kernel (polynomial, Gaussian, etc). Repeat steps (c) and (d).

2. Let $\mathrm{x} = \{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ be a subset of a input data set X. Consider a kernel function $k : \mathrm{X} \times \mathrm{X} \rightarrow \mathbb{R}$, which induces a feature space $\phi(\mathrm{X})$:

   (a) Deduce an expression, that allows to calculate the average distance to the center of mass of the image of set x in the feature space:

   $$\frac{1}{n} \sum_{i=1}^{n} \left\| \phi(\mathrm{x}_i) - \phi_S(\mathrm{x}) \right\|_{\phi(X)},$$

   where the center of mass is defined as

   $$\phi_S(\mathrm{x}) = \frac{1}{n} \sum_{i=1}^{n} \phi(\mathrm{x}_i).$$

   (b) Use previous expression to calculate the average distance to the center of mass of the following point set in $\mathbb{R}^2$, $\mathrm{x} = \{(0,1), (-1,3), (2,4), (3,-1), (-1,-2)\}$, in the feature spaces induced by the following kernels:

   i. $k(x,y) = \langle x, y \rangle$
   ii. $k(x,y) = \langle x, y \rangle^2$

iii. $k(x, y) = (\langle x, y \rangle + 1)^5$

iv. Gaussian kernel with $\sigma = 1$.

3. Controlling the model complexity

   (a) Download the Wisconsin Breast Cancer data set from `http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)` and divide it in a training set and a test set (50/50).

   (b) Train a SVM using a linear kernel. Find an optimal complexity parameter, C, plotting the training and test error vs. the complexity parameter. Use a logarithmic scale for C, $\left[2^{-5}, 2^{15}\right]$. Discuss the results.

   (c) Repeat item (b) using a Gaussian kernel with a fix σ value.

   (d) Repeat (c) varying σ and keeping C fixed.

4. Train an SVM for detecting whether a word belongs to English or Spanish:

   (a) Build a training and a test data sets. You can use the most frequent words in `http://en.wiktionary.org/wiki/Wiktionary:Frequency_lists`. Consider words at least 4 characters long and ignore accents.

   (b) Use an SVM software package that supports string kernels: LIBSVM, Shogun, etc.

   (c) Use cross validation to find an appropriate complexity parameter.

   (d) Evaluate the performance of the SVM in the test data set.