Assignment 3: Kernels and SVM's

Submission: Thursday October 1st
2 students per group

Prof. Fabio A. González
Machine Learning - 2015-II
Maestría en Ing. de Sistemas y Computación

---

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

   (a) Deduce an expression (using kernels) that, given a vector $w \in \mathrm{X}$, calculates the norm of the projection of the image of a point x, $\phi(x)$, onto the image of the vector w, $\phi(w)$:

   $$P_{\phi(w)}(\phi(x)) = \frac{\langle \phi(w), \phi(x) \rangle}{\|\phi(w)\|}$$

   (b) Deduce an expression (using kernels) to calculate the sample variance of the projections in the feature space of a set of points along a vector w:

   $$\mathrm{var}_{\phi(w)}(\mathrm{x}) = \frac{1}{n} \sum_{x_i \in \mathrm{x}} (P_{\phi(w)}(\phi(x_i)) - \mu)^2,$$

   where $\mu = \frac{1}{n} \sum_{x_i \in \mathrm{x}} P_{\phi(w)}(\phi(x_i))$.

   (c) Use the previous expression to calculate the variance of the projections of the images of the elements of the following point set in $\mathbb{R}^2$, $\mathrm{x} = \{(0,1), (-1,3), (2,4), (3,-1), (-1,-2)\}$ over the images of the vectors $w_1 = (1,1)$ and $w_2 = (-1,1)$, 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 $\sigma = 1$.

2. Regression on strings. Write an IPython notebook to document the following tasks:

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

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

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

   (b) Implement Kernel Ridge Regression (KRR).

   (c) Use the KRR implementation and the kernel k to train a model using the training data set in `http://fagonzalezo.github.io/ml/assign3-train.txt`. Evaluate the error of the model on the training data set. Plot the prediction of the model on the training data along with the real output values (results must be sorted descendently by the real output value).

(d) Evaluate the trained model on the test data set `http://fagonzalezo.github.io/ml/assign3-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 (b) and (c). For instance, the new kernel may be defined as:

$$k'(x, y) = (k(x, y) + 1)^d,$$

where d is positive integer exponent.

3. Digit recognition model understanding.

(a) Get the data from the Digit Recognizer problem `http://www.kaggle.com/c/digit-recognizer`.

(b) Choose two classes (e.g. 5 and 6, or 8 and 9) and train a linear SVM to discriminate between them. Find an optimal complexity parameter, C, plotting the training and test error vs. the regularization parameter. Use a logarithmic scale for C, $\{2^{-5}, 2^{-4}, \ldots, 2^{15}\}$. Discuss the results.

(c) Extract the weights of the classification model found in (b).

(d) Plot the discriminant function weights as follows:

i. Arrange the weights in a matrix with the same shape as the input image.

ii. Use a function such as `pcolor http://matplotlib.org/api/pyplot_api.html#matplotlib.pyplot.pcolor` to produce a color plot of the matrix.

iii. Use a diverging colormap that emphasizes negative and positive values `http://matplotlib.org/examples/color/colormaps_reference.html`.

iv. Discuss the results.

(e) Make a submission to Kaggle.

4. Train an SVM for detecting whether a word belongs to English or Spanish. Write an IPython notebook to document the following tasks:

(a) Build training and 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) Program a string kernel (it could be the one from the first problem).

(c) Use scikit-learn to train a SVM using a precomputed kernel.

(d) Use cross validation to find an appropriate regularization parameter.

(e) Evaluate the performance of the SVM in the test data set:

i. Build the confusion matrix.

ii. Illustrate examples of errors (English words mistaken as Spanish, Spanish words mistaken as English). Give a possible explanation for these mistakes.