Assignment 3: Kernels and SVMs

Submission: Monday April 2nd
2 students per group

Prof. Fabio A. González
Machine Learning - 2018-II
Maestría en Ing. de Sistemas y Computación

---

1. Let $\mathrm{x} = \{\mathrm{x}_1, \ldots, \mathrm{x}_n\}$ be a subset of a input data set X. Consider a kernel function $k : \mathrm{X} \times \mathrm{X} \to \mathbb{R}$, 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 (notice that the norm is **not** squared):

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

   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 the 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 $\sigma = 1$.

2. Digit recognition model understanding.

   (a) Get the data for the MNIST data set: `http://scikit-learn.org/stable/datasets/index.html#downloading-datasets-from-the-mldata-org-repository`.

   (b) Normalize your features so that each one has mean 0 and standard deviation 1.

   (c) Choose two classes (e.g. 1 and 0, or 6 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, $\left\{2^{-15}, 2^{-4}, \ldots, 2^{10}\right\}$. Discuss the results.

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

   (e) 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.

(f) Play with different pairs of digits and with different values for the C parameter (smaller values could produce smoother plots). Discuss the results.

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

   (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) Implement the kernel proposed by Lodhi et al. [1] as well as a kernel that counts the number of common n-grams between two strings. Alternatively use the implementation in `https://github.com/muggin/string-kernels`.

   (c) Use scikit-learn to train different SVMs using precomputed kernels. Use cross validation to find appropriate regularization parameters. Try different configurations of the parameters (λ and n).

   (d) Evaluate the performance of the SVMs in the test data set:

      i. Report the results in a table for the different evaluated configurations.

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

      iii. Discuss the results.

4. The assignment must be submitted as an Jupyter notebook through the following Dropbox file request, before midnight of the deadline date. The file must be named as `ml-assign3-unalusername1-unalusername2.ipynb`, where `unalusername` is the user name assigned by the university (include the usernames of all the members of the group). In case you need to include supporting files in addition to the notebook, submit a zipped file containing all the files and the notebook.

References

[1] Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., & Watkins, C. (2002). Text classification using string kernels. Journal of Machine Learning Research, 2(Feb), 419-444.