Assignment 3: Kernels and SVMs

Submission: Friday May 31st
2 students per group

Prof. Fabio A. González
Machine Learning - 2019-I
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, $\phi(w)$:

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

   (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:

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

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

   (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. 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, $\left\{2^{-15}, 2^{-14}, \ldots, 2^{10}\right\}$. 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 French 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 different string kernels:

i. **Histogram cosine kernel**: calculate a bag of n-grams representation (use the `CountVectorizer` from scikit-learn) and apply the `cosine_similarity` from scikit-learn.

ii. **Histogram intersection**: calculate a bag of n-grams representation, normalize it (the sum of the bins must be equal to 1) and calculate the sum of the minimum for each bin of the histogram.

iii. χ^2 **kernel**: calculate a bag of n-grams representation and apply the `chi2_kernel` from scikit-learn.

iv. **SSK kernel**: use the code available at this repository `https://github.com/helq/python-ssk`.

(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 (in particular different n values for the n-grams).

(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 (French words mistaken as Spanish, Spanish words mistaken as French). Give a possible explanation for these mistakes.

iii. Discuss the results.

4. The assignment must be submitted as a 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. Make sure that the notebook renders correctly and is free of errors before submitting.