Assignment 3: Kernels and SVMs

Submission: Sunday May 1th
3 students per group

Prof. Fabio A. González
Machine Learning - 2021-I
Maestría en Ing. de Sistemas y Computación

---

1. 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 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 $\forall i, ||x_i||_1 = 1.$) 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 plotting the training and validation error vs. the regularization parameter. Use a logarithmic scale for C, $\left\{2^{-15}, 2^{-14}, \ldots, 2^{10}\right\}$. 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 (English words mistaken as Spanish, Spanish words mistaken as English). Give a possible explanation for these mistakes.

      iii. Discuss the results.

2. SVM interpretability

   (a) Use the same dataset from question 1 and calculate a bag of n-grams representation.

   (b) Train a SVM using the histogram intersection kernel on this dataset.

   (c) Identify the support vectors found by the SVM training algorithm. Show the samples corresponding to the support vectors with the maximum absolute value of the dual coefficients ($\alpha_i y_i$), for both positive and negative values. Do they make sense? Analyze and discuss.

(d) For different test samples, calculate the classification manually, i.e. compute the kernel between the sample and the support vectors and check how they contribute, positively or negatively, to the final classification. Show those support vectors that have the highest value of the kernel. Analyze and discuss.

(e) Propose a method that for a given word to be classified, highlights in one color (e.g. blue) those characters that suggest the word is from English and in another color (e.g. red) those characters that suggest the word is from Spanish. Suggestion, for each n-gram in the vocabulary calculate its contribution to the classification; for each character in the word to classify, calculate its contribution to the classification.

3. Kernel logistic regression

We will implement a kernel version of logistic regression. The goal is to train a logistic regression model on a feature space F. Specifically, the discriminant function of the model is given by:

$$f(x) = P(C = 1|x) = \sigma(w\Phi(x)),$$

where $\Phi : X \to F$ is a mapping function associated with a kernel function $k : X \times X \to \mathbb{R}$ and σ is the logistic function.

Assume that the weight vector w is expressed as a linear combination of the training samples:

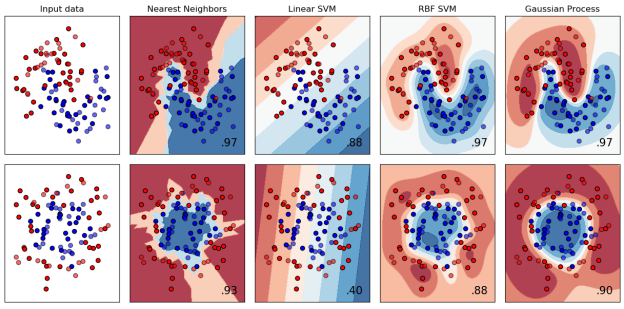
$$w = \sum_{i=1}^{\ell} \alpha_i \Phi(x_i)$$

(a) Write a expression of the discriminant function expressed in terms of the kernel and the coefficients α_i.

(b) Formulate the problem of learning the parameters of the model as an optimization problem that looks for the parameters α_i that minimize a cross entropy loss function.

(c) Write a function that receives a training data set and a kernel function and finds a vector α that minimizes the loss function using gradient descent.

(d) Test your algorithm using different kernels (linear, polynomial, Gaussian, etc.) on synthetic 2D datasets from sklearn (`https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html`). Plot the decision regions and 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-unalusername3.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.