Assignment 5

Machine Learning: Algorithms and Theory

Prof. Ulrike von Luxburg / Diego Fioravanti / Moritz Haas
Tobias Frangen / Siavash Haghiri

Summer term 2018 — due to **May 29th**

Readme

1. When you hand in your assignment you need to hand in the notebook too. Please do not write down a report with just results and/or figures. Ideally you should email it, it is easier to correct and more eco friendly, but we accept printed versions too. From now on not handing in the notebook will result in 0 points for the programming part.

2. Before the end of the course you need to present at least one of your solution in the tutorial. If you do not do that you cannot take the exam! If for any reason you cannot attend let us know, it is possible to change group or find another solution.

3. Join the class on ILIAS otherwise we cannot contact you if we need to

Exercise 1 (Logistic regression, 2+1+1+3 points) In this exercise you are going to use the dataset pretent in `candy.csv`. The dataset comes from `https://www.kaggle.com/fivethirtyeight/the-ultimate-halloween-candy-power-ranking/`

With it you are going to do two things, first learn how to take "real world" data and preprocess it and then apply logistic regression to find out if a candy contains chocolate or not.

(a) In the variable `candy` you can find the dataset. Now, divide the dataset in three parts: `names` that contains the names of the candies, `Y` that contains if a candy has chocolate or not and `X` everything else.

(b) Now, divide `X` and `Y` in `X_train`, `Y_train` and `X_test`, `Y_test`. The first $\frac{2}{3}$ (rounded down) of the dataset goes into train, the rest into test.

(c) Use scikit-learn `LogisticRegression` with the default paramenters to predict if a candy in `X_test` will contain chocolate or not. Compute the accuracy for your prediction. Remember, with accuracy more is better.

(d) `LogisticRegression` has a parameter C that can be tuned. Use `GridSearchCV` to do a 10-fold cross validation for $C \in \{0.01, 0.1, 1, 10, 100, 1000\}$. On the same graph, plot the cross validation accuracy and the test accuracy for the various C. Which C works best?

You can find the documentation for `GridSearchCV` here `http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html` in the results the key that you are interested in is `mean_test_score`.

`LogisticRegression` has a method called `score(X, Y)` that computes the accuracy of the prediction. You cannot use it for c) but you can use it here See the documentation for details `http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

Exercise 2 (Linear support vector machines (SVM), 3+2+2+1 points)
In this exercise you will apply linear SVM to understand its behavior. And you will learn how to plot decision boundaries with matplotlib.

(a) First we consider a dataset $(X_i, Y_i)_{i=1..n}$. The first 300 points will be our train set and the rest the test one. Plot the train data set. Then apply linear SVM using `LinearSVC`. The decision boundary associated with the model is a hyperplane, in this case a line, separating the two classes. Add the decision boundary to the plotted data. Use as much code as you like from `http://scikit-learn.org/stable/auto_examples/svm/plot_iris.html` and read the explanation in the beginning.

(b) The regularization parameter C is by default equal to 1 and can be changed by specifying `LinearSVC(C=3)`. Vary $C \in \{0.001, 1, 1000\}$ and plot the new hyperplane in order to see its influence. How does the hyperplane change with increasing C, does it become more or less data dependent? Calculate the training errors for each C and plot it.

(c) As in 1d) use `GridSearchCV` and perform cross validation to determine the best $C \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$. On the same graph, plot the cross validation accuracy and the test accuracy for the various C. Which C works best?

(d) In `X_1`, `Y_1` and `X_2`, `Y_2` you find two datasets generated in a slightly different way. Run a linear SVM on them and compute the accuracy using `clf.score`. Can you explain what you observe ? After doing that, plot the datasets. Do the plots align with your previous explanation?

Exercise 3 (Primal hard margin SVM problem, 3+4 points) Given training data $(X_i, Y_i)_{i=1,\ldots,n} \in (\mathbb{R}^d \times \{-1,+1\})^n$ the primal hard margin SVM problem is given as

$$\begin{aligned} &\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|w\|^2 \\ &\text{subject to } Y_i\left(w^T X_i + b\right) \geq 1, \forall\, i = 1, \ldots, n. \end{aligned} \tag{1}$$

(a) Recall the meaning of a hyperplane in canonical representation. Show that any solution of (1) gives rise to a hyperplane in canonical representation.

(b) Assume the training data is linearly separable, that is there exists a solution of (1). Show that this solution is unique.