Assignment 4

Statistical Machine Learning

Moritz Haas / Prof. Ulrike von Luxburg

Summer term 2022 — due on **May 16th at 12:00**

Exercise 1 (Logistic regression, Cross Validation and Performance Measures, 2+2+3+3+2+1=13 points)

In this exercise we are going to use the candy dataset from kaggle. It is available from our website as `candy-data.csv`. With this dataset we are going to do two things. First we will apply logistic regression to find out if a candy contains chocolate or not. Second, we will study the effects that regularization has on model training and predictions under different evaluation criteria.

We are also going to use the popular machine learning framework `scikit-learn`. Instead of implementing the machine learning algorithms ourselves, as we did for kNN and ridge regression in the last two assignments, we are now going to use the implementation of logistic regression that comes with `scikit-learn`. We will continue to use this framework in future assignments.

(a) In the variable `candies` you can find the dataset that we loaded for you. Now, divide the dataset in three parts: `names` that contains the names of the candies, `ys` that contains if a candy has chocolate or not and `xs` everything else. Now, divide `xs` and `ys` in `xs_train`, `ys_train` and `xs_test`, `ys_test`. The first $\frac{2}{3}$ (rounded down) of the dataset goes into train, the rest into test.

(b) Use scikit-learn `LogisticRegression` with the solver `'liblinear'` and otherwise default parameters to predict if a candy in `xs_test` contains chocolate or not. Compute the accuracy for your prediction and store it in a variable `acc`.

(c) Use scikit-learn `GridSearchCV` to do a 10-fold cross validation for $C \in \{0.01, 0.1, 1, 10, 100, 1000\}$. Plot on the same graph the cross validation error and the test error. Please use a log scale when plotting C. Does cross validation choose the best C? Comment in terms of over/underfitting.

Hint: For the CV error, the key that you are interested in is `mean_test_score`.

`LogisticRegression` has a method called `score(X, Y)` that computes the accuracy between the predictions and the true `Y`. You cannot use it for b) but you can use it here.

(d) For every C compute the recall, ROC curve and AUC on the test set. Which hyperparameters perform best in the respective performance measure?

Hint: Use the functions `roc_curve`, `roc_auc_score` from `sklearn.metrics`.

(e) Plot the ROC-curve on the test set for the values of C, that perform best in some performance measure. Name 3 desired properties of a classifier; for each of these decide which value of C you would choose.

(f) Which cardinal crime have we just committed in (e)?

Exercise 2 (Probabilistic Point of View on Ridge Regression, 1+1+2=4)

Here we consider the linear regression setting,

$$Y = Xw + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2 I)$, $X \in \mathbb{R}^{n \times d}$, $Y \in \mathbb{R}^n$, $w \in \mathbb{R}^d$ and $\varepsilon \in \mathbb{R}^n$.

(a) Determine the distribution of $Y|X, w$.

(b) Now assume any prior distribution on the weights and independence between X and w. Rewrite the likelihood $P(w|X, Y)$ in terms of the posteriors $P(Y|X, w)$ and $P(Y|X)$, and the prior $P(w)$. It should become clear where the independence assumption is used.

(c) Now assume $w \sim N(0, \tau^2 I)$. Show that the maximum likelihood solution w^* corresponds to the solution of Ridge regression. What is the value of the regularisation parameter λ in terms of σ^2 and τ^2.

Exercise 3 (Feedback Poll, 2 bonus points)
Every week there will be a feedback survey on Ilias about the lecture, tutorials and exercise sheets. If you participate, it helps us in understanding what we can improve.