WBI - Machine Learning
Dr. Patrick Jähnichen

Basics of Machine Learning, Winter Term 2017/2018

Exercise sheet 4

Discussion on November 28st and December 1st, 2017

Task 1: Theory

We have learned about ridge regression, where we add a penalty on large values for parameters. This is also known as *shrinkage penalty*, as it lets the parameter values shrink to 0. Recall the ridge regression model (in the multidimensional case) with the objective function

$$J(\theta) = (Y - X\theta)^\top (Y - X\theta) + \delta^2 \theta^\top \theta.$$

We have so far looked at this in terms of the least squares fit, i.e. in terms of minimizing the sum of squared errors. Applying the likelihood principle, we assume a normally distributed noise on our data, inducing some distribution. In the above case we assume that

$$Y_i \sim \mathcal{N}(\theta^\top X_i, \sigma^2)$$

for the data.

(a) Inspect the second part of the cost function above. What distributional assumption are we placing on θ?

(b) Construct the according likelihood function for ridge regression.

(c) Make the same distributional assumption on θ in case of logistic regression and construct the lieklihood function.

(d) Derive the gradient of the log-likelihood w.r.t. θ when making that assumption.

Task 2: Practical

Download the Spambase data set[1]. Your task is to classify emails into spam or not-spam (sometimes called ham) emails. There are of course sophisticated methods to do this, but we will look into logistic regression again.

(a) Use a suitable library to load the data. Compile a training set and a test set from the data. A description of the available features (input dimensions) is given in spambase.names. Most of the data are in the $[0, 100]$ interval but some are not normalized. More info in the spambase.names file.

(b) Use your logistic regression implementation from the last exercise and report the F_β measure for $\beta = \{.5, 1\}$. This is given by

$$F_\beta = (1 + \beta^2) \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}.$$

The precision is given as the fraction of correctly classified spam mails (*true postives*) and the sum of all mails classified as spam, including incorrect classifications (*true positives + false positives*). In contrast, recall is defined as the fraction of correctly classified spam mails (*true positives*) and all spam mails (*true positives + false negatives*).

(c) Implement a logistic ridge regression algorithm. Determine the best parameter δ^2 by 5-fold cross validation. Plot parameter values against δ^2 and report your results on the test set as above.

Important: the test set created above is **not** the validation set!

[1] `https://archive.ics.uci.edu/ml/datasets/Spambase`