# Homework 4, Machine Learning

## Linear regression and logistic regression

## 1 Linear regression

In the first part of this homework the task is to implement linear regression model and experiment on a given dataset. Linear regression can be implemented using normal equations:

$$\boldsymbol{\theta} = (X^T X)^{-1} X^T \mathbf{y}, \tag{1}$$

where $X$ is the input matrix and $\mathbf{y}$ is the column vector of labels. With $\ell_2$-regularization the normal equation will get the form:

$$\boldsymbol{\theta} = (X^T X + \lambda I)^{-1} X^T \mathbf{y}, \tag{2}$$

where $\lambda$ is the regularization parameter.

### 1.1 Data

The data contains the following features: sex, systolic blood bressure, diastolic blood pressure, cholesterol level, age and body mass index. The first column provides the binary label whether the systolic blood pressure is high or not. The goal in this first task is to predict the systolic blood pressure value from the other features. You have to omit the first column with the binary label and treat the second column as the predicted value. Don't forget to add a column of ones as the first column to be used as dummy features for the intercept term.

### 1.2 Experiments

Learn the regressor function that best predicts the systolic blood pressure. For doing that, several aspects can be varied: the set of features to use (easiest is to use just the dbp), polynomial features (squared, cubed, etc), regularization with different values of regularization parameter.

### 1.3 Evaluation

For evaluation use the k-fold cross validation. You can compute the squared loss on the validation set and use it for comparison:

$$cost = \sum_{i=1}^{N_{val}} (\boldsymbol{\theta}^T \mathbf{x}_i - y_i)^2 \tag{3}$$

### 1.4 Write-up

The report should include a short description of the task, data and the implementation (including how to run it and what options can be set). It should contain the detailed description of the experiments you did, what parameters did you vary, how the results changed with different parameter values.

When you vary some parameter value and record the squared loss with each value, it would be good to represent such results with a figure that plots the squared loss as a function of the parameter.

Also, add plots that represent on the same figure the real data and the predictions. Although you may have more features than just one, on the figure you can plot the real sbp and the predicted sbp as a function of just one feature (for example dbp), even when the real feature vector includes more data than just this one feature.

The report should state also clearly, which setting according to your experiments produces the best results.

### 1.5 Toolkits

You can also use some toolkit or library to write the wrapper for executing experiments. For python the recommended toolkit is `scikit-learn`: `http://scikit-learn.org`. Look for examples in `http://scikit-learn.org/stable/modules/linear_model.html` and the references in `http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html` and `http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html`.

## 2 Logistic regression

The second task in this homework is to implement the logistic regression classification model using Newton's method. Newton's method updates the parameters iteratively using both the first and second partial derivatives:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - H^{-1} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^{(k)}) \tag{4}$$

The components in the vector of the first partial derivative $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ can be computed as:

$$\frac{\partial \ell(\boldsymbol{\theta})}{\theta_j} = \sum_{i=1}^{m} (y_i - h_{\boldsymbol{\theta}}(\mathbf{x}_i)) x_{ij} \tag{5}$$

The components of the Hessian (second partial derivatives) are computed as:

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} = -\sum_{i=1}^{m} h_{\boldsymbol{\theta}}(\mathbf{x}_i)(1 - h_{\boldsymbol{\theta}}(\mathbf{x}_i))x_{ij}, x_{ik} \tag{6}$$

## 2.1 Data

The data contains the following features: sex, systolic blood pressure, diastolic blood pressure, cholesterol level, age and body mass index. The label is in the first column. The goal is to predict whether the person has high systolic blood pressure or not. Therefore you should exclude the column with sbp values from training. The data is awailable from the course web page. Don't forget to add a column of ones as the first column to be used as dummy features for the intercept term.

## 2.2 Experiments

Learn the logistic regression model that best predicts the two classes. For doing that, several aspects can be varied: the set of features to use, polynomial features (squared, cubed, etc), regularization with different values of regularization parameter.

## 2.3 Evaluation

For evaluation use K-fold cross-validation to compute the classification accuracy. Report the average accuracy of all the folds.

## 2.4 Write-up

The report should include a short description of the task, data and the implementation (including how to run it and what options can be set). It should contain the detailed description of the experiments you did, what parameters did you vary, how the results changed with different parameter values.

When you vary some parameter value and record the squared loss with each value, it would be good to represent such results with a figure that plots the cross-validation accuracy as a function of the parameter.

The report should state also clearly, which setting according to your experiments produces the best results.

## 2.5 Toolkits

You can also use some toolkit or library to write the wrapper for executing experiments. For python the recommended toolkit is `scikit-learn`: `http://scikit-learn.org`. Look for the reference in `http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`.