Pattern Recognition: Assignment 7

Due on Tuesday, June 19 2012, 23:59

Bernhard X. Kausler (TA), Summer Term 2012

patternrecognition@hci.iwr.uni-heidelberg.de

`http://hci.iwr.uni-heidelberg.de/MIP/Teaching/pr/`

Regression

In this assignment you will learn about linear and logistic regression. You will write your own implementation of linear regression and will investigate the class posterior density and confidence limits of the decision boundary obtained with logistic regression.

Introduction to R

If you are new to R, please familiarize yourself with the most important commands (see `R_intro.R` which is a one-to-one translation of the `matlab_intro.m` to R to simplify the transition and show similarities). Feel free to skip this step, if you are a proficient R user.

Prob. 1: Linear Regression

In linear regression we are fitting the model

$$h_\theta(x) = \theta^T x$$

by minimizing the error J:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

(a) Implement linear regression (3 points)

In the lecture we learned, that we can obtain the optimal parameters θ using the closed-form expression

$$\theta = (X^T X)^{-1} X^T \vec{y}$$

Implement linear regression as a R function using this expression.

(b) Apply your linear regression implementation(2 points)

The file `cities.csv` contains samples describing the profit of a pizza delivery chain depending on the size of a city's population (both in multiples of 10.000). Use your implementation of linear regression to fit a linear model with `population` as regressor and `profit` as response. Make a scatterplot of the data and add the linear model as a line to the plot.

(c) Visualize the error function (2 points)

Firstly, make a 3d surface plot of the error function J depending on the two parameters θ_1 and θ_2. Use the city-profit data for $\vec{x}$ and $\vec{y}$.
Secondly, make a contour plot of the error function and add the optimal parameters you found in section (a) as a point to the plot. Where does the point lie?
How many minima of the error function do you see? Why?

Prob. 2: Logistic Regression

In this problem we apply logistic regression and visualize the results. Fig. 1 shows a small clipping of the plot you are supposed to produce.

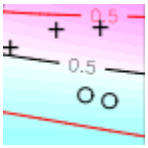

Figure 1: Logistic regression: color-coded class posterior probabilities together with a decision boundary and 95% confidence limits

Logistic regression models the posterior density of a class C_1 as

$$p(C_1|x) = y(x) = \sigma(\theta^T x)$$

with $p(C_2|x) = 1 - p(C_1|x)$. The function

$$\sigma(\theta^T x) = \frac{1}{1 + e^{-(\theta^T x)}}$$

is called the logistic sigmoid.

(a) Derivative of the logistic sigmoid (2 points)

Verify the derivate of the logistic sigmoid $\frac{d\sigma}{da} = \sigma(1 - \sigma)$. Where do you need this equation in logistic regression?

(b) Apply logistic regression(4 points)

Train a logistic regression model on the students' score data. Use the two scores as regressors and exam.passed as response. To visualize the result make a scatterplot of the data and add the class posterior probabilities obtained from the model as a filled contour plot.
What functional form does the decision boundary have? Give a closed-form expression and add the boundary to the plot. (suggest command: glm, filled.contour, abline)

(c) Polynomial features (3 points)

Explicitly project to a higher dimensional space by adding 2nd degree polynomials of the two `score` regressors: score1^2, $\text{score1} \cdot \text{score2}$, and score2^2
Again, train a logistic regression model on the extended feature space and plot the data together with the class posteriors as a filled contour and the decision boundary as a line. (This time you don't have to give a closed-form expression for the decision boundary.)
How does the decision boundary look like? Compare with the decision boundary you obtained in section (a). Explain the effect the polynomial features on the decision boundary. (suggested command: contour)

(d) Confidence limits (4 points)

Determine the 95% confidence interval of the class posterior density. Use the standard error as given by the R function `predict`.
Add the upper and lower confidence limits as red lines to the plot of section (b) (see Fig. 1 for an example). (suggest command: `predict` with paramter `se.fit=TRUE`)

Regulations

Please hand in the matlab code, figures and explanations (describing clearly which belongs to which). Non-trivial sections of your code should be explained with short comments, and variables should have self-explanatory names. Plots should have informative axis labels, legends and captions. Please enclose everything into a single PDF document (e.g. use the publish command of MATLAB for creating a LaTeX document and run latex, dvips and ps2pdf or copy and paste everything into an office document and convert to PDF). Please email the PDF to patternrecognition@hci.iwr.uni-heidelberg.de before the deadline specified below. You may hand in the exercises in teams of two people, which must be clearly named on the solution sheet (one email is sufficient). Discussions between different teams about the exercises are encouraged, but the code must not be copied verbatim (the same holds for any implementations which may be available on the WWW). Please respect particularly this rule, otherwise we cannot give you a passing grade.