Exercises for
Pattern Recognition
Sebastian Käppler, Nooshin Haji Ghassemi
Assignment 30.11/, 1.12.2015

General Information:

Exercises (1 SWS):	Mo 12:15 – 13:30 (H10 lecture hall building) and Tue 08:45 – 10 (0.151-113)
Certificate:	Oral exam at the end of the semester
Contact:	sebastian.kaeppler@fau.de, nooshin.haji@fau.de

Regression

Exercise 1 The goal of this exercise is robust regression line fitting for N measurements (x_i, y_i). Thus, you should estimate parameters a, b for a line $ax_i + b$ that best explains your observations y_i. Here we employ the Huber norm to make the estimate more robust to outliers compared to simple least-square regression:

$$(a, b) = \arg\min_{a,b} D(a, b) = \arg\min_{a,b} \sum_{i=1}^{N} \phi_{\text{Huber}}(y_i - ax_i - b) \tag{1}$$

The parameters (a, b) are determined using iterative numerical optimization. The Huber norm is defined as

$$\phi_{\text{Huber}}(z) = \begin{cases} z^2 & \text{if } |z| \leq M \\ M(2|z| - M) & \text{if } |z| > M \end{cases} \tag{2}$$

(a) Calculate the gradient of the cost function w.r.t. a and b. The gradient is necessary for many iterative numerical optimization techniques.
Hint: You need to calculate the derivative of the Huber norm.

(b) Show that the Huber norm is convex. Use the first-order convexity condition for differentiable functions $f(x)$

$$f(z) \geq f(x) + f'(x)(z - x)$$

Start by proving convexity for $g(x) = x^2$ and $h(x) = M(2|x| - M)$. Then, treat the special cases that occur due to the piece-wise definition of the Huber norm. For this exercise, focus only on positive values x, z, M.

(c) Download the provided measurements from the exercise homepage. Minimize the Huber norm using MATLAB. You do not need the Classification Toolbox. Use the MATLAB function `fminunc`.

(d) Compare the robust line fitting to a ordinary least-square approach. Find situations where the robust approach is superior. Show that due to convexity, the optimum is always found.

Exercise 2 A training set of N independent samples with feature vectors $\boldsymbol{a}_i \in \mathbb{R}^D$ and target variables $b_i \in \mathbb{R}$ is given. A linear model with the parameter $\boldsymbol{x} \in \mathbb{R}^D$ is assumed to estimate the target variable from the feature $b = \boldsymbol{x}^T \boldsymbol{a}$.

Ridge regression is least-squares linear regression with L_2-norm regularization. It is defined by the optimization problem

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmin}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \lambda \|\boldsymbol{x}\|_2^2 \quad , \tag{3}$$

with the design matrix $\boldsymbol{A} \in \mathbb{R}^{N \times D}, \boldsymbol{A}(i,j) = \boldsymbol{a}_i(j)$ and the target vector $\boldsymbol{b} \in \mathbb{R}^D, \boldsymbol{b}(i) = b_i$.

(a) Derive the solution of the ridge regression optimization problem.

(b) What is the effect of the regularization?

(c) Ridge regression can be motivated by Maximum A Posteriori (MAP) estimation. In MAP estimation, the a posteriori probability of the parameters after observing the training data is maximized $\boldsymbol{x}^* = \operatorname{argmax}_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{A}, \boldsymbol{b})$. The assumption of Gaussian noise $p(b|\boldsymbol{x}, \boldsymbol{a}) = \mathcal{N}\left(b|\boldsymbol{x}^T\boldsymbol{a}, \beta^{-1}\right)$ and a Gaussian prior for the parameters $p(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{x}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}\right)$ is made. Show that MAP estimation in this setting is equivalent to ridge regression.