Machine Learning 2007
(Home work 1)

May 25, 2007

- Due on Wednesday, June 13, before 10 a.m.
- Late submissions will not be accepted.
- Submit hard copy of the results, plots and your workings
- Submit a printed copy of the codes also.
- You may save time if you use MATLAB for the computations and plots.
- Please do not hesitate to contact me if you do not understand the problems.

1. **[10 points + 10 points + 10 points] Regression**

   (a) Given the real-valued function $f(x_1, x_2, ..., x_n)$, if all partial second derivatives of f exist, then the Hessian matrix of f is the matrix

   $$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

   Let $J(\boldsymbol{\theta})$ be the usual cost function for linear regression for a data in $\Re$, i.e.,

   $$J(\theta_0, \theta_1) = \sum_{i=1}^{n} ((\theta_0 + \theta_1 x^{(i)}) - y^{(i)})^2.$$

   Compute the Hessian matrix H for $J(\boldsymbol{\theta})$. Show that H is positive definite, i.e., for any $\boldsymbol{z} \in \Re^2$, $\boldsymbol{z}^T H \boldsymbol{z} \geq 0$.

   Note: This shows that that the cost function for linear regression is convex, and it has no local minimas but a single global minima. This result is true for higher dimensions also You may try proving it, but no points for it.

(b) Download the data data.dat from the course website. Use the online gradient descent rule to fit a line on the data. Plot the points along with the line. Give the equation of the fitted line.

(c) Download the data class.txt from the course website. Plot the points, you should use a different marker for denoting points in the two different classes. Now, plot the decision boundary obtained by logistic regression. Give the equation of the decision boundary.

2. **[10 +15 +10 +10 +5 points] Bayesian Decision Theory**

(a) A patient either has a certain form of cancer or not. A biopsy will return either $\oplus$ meaning that the patient is sick, or $\ominus$. However, the biopsy only has 98% accuracy in identifying $\oplus$ and a 97% accuracy in identifying $\ominus$. Also, we know that the prior probability that a random person has this disease is 0.008. What is the probability that a person for whom the test returns $\oplus$ has the disease?

(b) Let x have an exponential density:

$$p(x|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Suppose n samples $x_1, x_2, \ldots, x_n$ are drawn independently according to $p(x|\theta)$. Show that the maximum likelihood estimate for θ is given by

$$\hat{\theta} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}$$

(c) Consider a two class classification problem with two classes c_1 and c_2 with the following prior and class conditional distributions: $P(c_1) = P(c_2) = 0.5$, $p(\boldsymbol{x}|c_1) = \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ and $p(\boldsymbol{x}|c_2) = \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$. Derive the equation of the Bayesian discriminant functions for the following values of μ and Σ. Plot the decision boundaries.

- $\boldsymbol{\mu}_1 = [2, 8]^T$, $\boldsymbol{\mu}_2 = [8, 2]^T$, $\Sigma_1 = \Sigma_2 = \begin{bmatrix} 3.0 & 0.0 \\ 0.0 & 3.0 \end{bmatrix}$
- $\boldsymbol{\mu}_1 = [3, 6]^T$, $\boldsymbol{\mu}_2 = [3, -3]^T$, $\Sigma_1 = \begin{bmatrix} 0.5 & 0.0 \\ 0.0 & 2.5 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 2.0 & 0.0 \\ 0.0 & 2.0 \end{bmatrix}$
- $\boldsymbol{\mu}_1 = [3, 6]^T$, $\boldsymbol{\mu}_2 = [3, -3]^T$, $\Sigma_1 = \begin{bmatrix} 0.5 & 0.5 \\ 0.0 & 2.5 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 2.0 & 0.5 \\ 0.0 & 2.0 \end{bmatrix}$

(d) Download the file class2.txt from the course website. It has data for a two class classification problem. The features are in $\Re^2$ and the classes are denoted by 0 and 1. Assume that the data from the two classes are generated from a normal distribution. Estimate the prior probabilities and the class conditinal densities for the two classes.

(e) Build a Bayes Classifier using the probability values obtained from the previous problem. Now download the file class3.txt. Ignore the labels in the data and classify the data points using the classifier. Use the class labels to compute the number of misclassifications done by your classifier.

3. **[10 + 15 points] Non parametric methods**

   (a) Download the file trgIris.dat from the course website. This file contains a classification data, it has four features and 3 classes. Use the first 100 data points in the file as a training set and classify the rest 50 points using k-NN rule. Report the number of misclassifications obtained by using $k = 3, 5, 7$ and 9.

   (b) Recall weighted linear regression discussed in class. For a regression problem we want to weight the cost function for different query points. We use a linear function as the local model. Specifically we define a cost function as

$$J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n} w_i(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} - y^{(i)})$$

   where

$$w_i = \exp - \left(\frac{||\boldsymbol{x} - \boldsymbol{x}^{(i)}||^2}{2\tau^2}\right).$$

   Using this cost function implement a locally weighted regression method to fit a function to the data in wr.dat. (Find out the predicted values for the following x values: -5,-4,-3,-2,-1,0,1,2,3,4,5,6,7,8,9,10,11,12, and plot the function using these values only). Use $\tau = 0.1, 3.0, 8.0$ and 10.0 and draw four plots of the fitted function for the different values of τ. Comment briefly on what happens to the fit when τ is too small or too large.