CS4801: Principle of Machine Learning
Assignment 1

Due on 27th August

This homework consists of problems covering, linear algebra, probability, MLE and introduction to Machine Learning. A few instructions to make life easier for all of us:

- Assignment need not to be submitted.
- We will have a short quiz on the question of this assignment (5 points) for 10 minutes on 28th August 08:00 am.
- In short quiz please write concisely and clearly. There are points for intermediate steps, but not in “talking problems to death.”

1 Conceptual Exercises

Exercise 1 : Application

(a) (1.5 points) Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in prediction. Finally, provide n(number of sample) and p(number of features).

i. We collect a set of data on the top 500 companies in India. For each company we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

ii. We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

iii. We are interested in predicting the % of change in the INR in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the INR, the % change in the US market, the % change in the British market, and the % change in the German market.

Exercise 2: Linear Algebra

(a) (3 points) Any real, symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ has the decomposition $A = \mathbf{U}\Sigma\mathbf{U}^T$, where $\Sigma \in \mathbb{R}^{n \times n}$ is a diagonal matrix with the eigenvalues $\lambda_1, \ldots, \lambda_n$ of $\mathbf{A}$ on the diagonal and $\mathbf{U}$ is an orthogonal matrix in $\mathbb{R}^{n \times n}$, that is $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}$, which contains the corresponding eigenvectors (more precisely: an orthogonal basis of the eigenspace of the corresponding eigenvalue).

i. What are the eigenvalues and eigenvectors of $\mathbf{A}^k$ (matrix product with itself) for $k \in \mathbb{N}$.

ii. Let $\mathbf{U} \in \mathbb{R}^{n \times n}$ be an orthogonal matrix - show that $\|\mathbf{U}^T\mathbf{x}\|_2 = 1$ for any vector $\mathbf{x} \in \mathbb{R}^n$ with $\|\mathbf{x}\|_2 = 1$.

iii. Use the previous result to show that $max_{\|\mathbf{x}\|_2=1}\mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda_{max}$, where λ_{max} is the largest eigenvalue of $\mathbf{A}$. [hint: $max_{\|\mathbf{x}\|_2=1}\mathbf{x}^T\mathbf{A}\mathbf{x}$ means $max\mathbf{x}^T\mathbf{A}\mathbf{x}$ such that $\|\mathbf{x}\|_2 = 1$ (a constraint)]

[Hint: The last part can be solved using the decomposition of $\mathbf{A}$ and then doing a variable transformation $\mathbf{y} = \mathbf{U}^T\mathbf{x}$. Then one gets a very simple optimization problem for $\mathbf{y}$.]

Exercise 3: Multivariate Analysis

(a) (3 points) Compute the derivative $\nabla_{\mathbf{x}} f$ for following functions $f : \mathbb{R}^d \to \mathbb{R}$ with respect to $\mathbf{x}$, where $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{d \times d}$, $B = (b_{ij}) \in \mathbb{R}^{m \times d}$

i. $f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle = \mathbf{x}^T\mathbf{w} \qquad \Rightarrow \ \nabla_{\mathbf{x}} f = \mathbf{w}$,

ii. $f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{A}\mathbf{x} \rangle = \mathbf{x}^T\mathbf{A}\mathbf{x} \qquad \Rightarrow \ \nabla_{\mathbf{x}} f = \mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x}$,

iii. $f(\mathbf{x}) = \|\mathbf{B}\mathbf{x}\|_2^2 \qquad \Rightarrow \ \nabla_{\mathbf{x}} f = 2\mathbf{B}^T\mathbf{B}\mathbf{x}$,

Exercise 4: Basic Probability

(a) (1 point) X and Y are two random variables and $Y = mX + c$ where m and c are not random variables. Prove that the correlation coefficient of X and Y is 1. [Hints : correlation coefficient is $corr(X, Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$]

(b) (2+1 points) Given a set of samples $\mathcal{D} = \{(\mathbf{x}_i)_{i=1}^N\}$ where $\mathbf{x} \in \mathbb{R}^d$, derive the maximum likelihood estimation for parameter of following distributions

i. $\mathbf{x}$ be a d-dimensional real vector with multi-variate Gaussian distribution, i.e.,

$$P(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$$

where $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ are parameters (mean and covariance) of the distribution. Covariance matrix is symmetric and positive semi-definite.

- Find the maximum likelihood estimation for μ.
- Find the maximum likelihood estimation for Σ.

ii. $\mathbf{x}$ be a d-dimensional binary (0 or 1) vector with a multinoulli distribution or categorical distribution

$$P(\mathbf{x}|\mu) = \prod_{k=1}^{d} \mu_k^{x_k},$$

where $\mu \in \mathbb{R}^d$ is parameter vector, μ_k being the probability of $x_k = 1$ and hence $\sum_{k=1}^d \mu_k = 1$. Find the maximum likelihood estimation for μ.

(c) (1 point)$X \in \mathbb{R}^+$ is a non-negative random variable. Prove that, for any positive real number $a > 0$

$$P(X \geq a) \leq \frac{E[X]}{a},$$

where $E[X]$ is expectation of X.

(d) (5 points)$X \in \mathbb{R}$ and $Y \in \mathbb{R}$ are two random variables. Given a parameter w, the conditional distribution of Y is given as

$$p(Y|w, \sigma, X = x) \sim \mathbb{N}(wx, \sigma),$$

while σ is known. Find out maximum likelihood estimation of w from set of samples $\mathcal{D} = \{(x_i, y_i)_{i=1}^{N}\}$.

(e) (5 points) It is a continuation of previous problem. Given $w \sim \mathbb{N}(0, \beta)$, find out maximum a posteriori (MAP) estimation of w from set of samples $\mathcal{D} = \{(x_i, y_i)_{i=1}^{N}\}$. MAP implies the value of w for which the posterior density function of w, i.e. $p(w|\mathcal{D})$ is maximum.[hint: use Bayes theorem]

(f) (5 points) $X \in \mathbb{R}$ is a continuous random variable and $Y \sim Bernoulli(0.5)$. Conditional probability distribution of X is given by

$$p(X|Y = y) \sim \mathbb{N}(\mu_y, \sigma^2).$$

Note that, means for two conditional distributions are different (μ_0 and μ_1) but the variance is same σ^2. Using Bayes theorem, compute the ratio $\frac{P(Y=1|X=x)}{P(Y=0|X=x)}$.