Introduction to Machine Learning **Fall Semester, 2016/17**

Homework 0: Nov 2nd, 2016

Due: Nov 16th (See the submission guidelines in the course web site)

Linear Algebra

1. Let $\mathbf{v_1}, \ldots, \mathbf{v_p} \in \mathbb{R}^n$ be orthonormal vectors ($p < n$). Let:

$$\mathbf{x} = c_1\mathbf{v_1} + \cdots + c_p\mathbf{v_p}$$

   Show that:

$$\|x\|^2 = c_1^2 + \cdots + c_p^2$$

2. Let A be a matrix over $\mathbb{R}$ and let λ be an eigenvalue of A.
   (a) Show that λ^2 is an eigenvalue of A^2.
   (b) For a polynomial $p(x) = c_0 + c_1x + c_2x^2 + \ldots + c_kx^k$ (with $c_i \in \mathbb{R}$), define

$$p(A) = c_0I + c_1A + c_2A^2 + \ldots + c_kA^k$$

   Show that $p(\lambda)$ is an eigenvalue of $p(A)$.

3. Reminder: If A is a $n \times n$ matrix, its *trace* is defined as $tr(A) = a_{11} + \cdots + a_{nn}$. Assume A is diagonalizable with eigenvalues $\lambda_1, \ldots, \lambda_n$.
   (a) Show that $det(A) = \lambda_1 \cdot \ldots \cdot \lambda_n$.
   (b) Let B also be a $n \times n$ matrix. Show that $tr(AB) = tr(BA)$.
   (c) Show that $tr(A) = \lambda_1 + \ldots + \lambda_n$.

4. A matrix A over $\mathbb{R}$ is called *positive semidefinite* if for every vector v, $v^\mathsf{T}Av \geq 0$. Show that the set of all symmetric positive semidefinite matrices is *convex*: Namely, that for any two symmetric positive semidefinite matrices A, B and a scalar $0 \leq \theta \leq 1$, $\theta A + (1-\theta)B$ is also symmetric positive semidefinite.

Calculus and Probability

1. Reminder: $X \sim Pois(\lambda)$ if for $k = 0, 1, 2, \ldots$, $Pr(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$. Show that if $X \sim Pois(\lambda)$, $\mathrm{E}(X) = \lambda$.

2. Let X, Y be two random variables defined over the same probability space. Recall that the *conditional expectation* $\mathrm{E}(X|Y)$ is also a random variable.
   (a) Show:

$$\mathrm{E}(X) = \mathrm{E}_Y(\mathrm{E}(X|Y))$$

   (b) (Bonus:) Assume the variance of Y is finite. Show:

$$\mathrm{Var}[Y] = \mathrm{E}(\mathrm{Var}[Y|X]) + \mathrm{Var}[\mathrm{E}(Y|X)]$$

3. *Matrix calculus* is the extension of notions from calculus to matrices and vectors. We define the derivative of a scalar y with respect to a vector $\mathbf{x}$ as the column vector which obeys:

$$\left(\frac{\partial y}{\partial \mathbf{x}}\right)_i = \frac{\partial y}{\partial x_i}$$

for $i = 1, \ldots, n$. Let A be a $n \times n$ matrix. Prove that:

$$\frac{\partial \mathbf{x}^T A \mathbf{x}}{\partial \mathbf{x}} = (A + A^T)\mathbf{x}$$

4. (a) Use the method of Lagrange multipliers to find the maximum of $f(x, y) = x^3 y^5$ with respect to the constraint $x + y = 8$.

   (b) Let $\mathbf{p} = (p_1, \ldots, p_n)$ be a discrete distribution, with $\sum_{i=1}^n p_i = 1$ and $p_i \geq 0$ for $i = 1, \ldots, n$. The *entropy*, which measures the uncertainty of a distribution, is defined by:

$$H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log p_i$$

   where we define $0 \log 0 = 0$. Use Lagrange multipliers to show that the uniform distribution has the largest entropy (Tip: Ignore the inequality constraints $p_i \geq 0$ and show that the solution satisfies them regardless).

   Bonus: Can you give a simpler argument?

Decision Rules and Concentration Bounds

1. Let $\mathbf{X} = (X_1, \ldots, X_n)^\mathsf{T}$ be a vector of random variables. $\mathbf{X}$ is said to have a **multivariate normal (or Gaussian) distribution** with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and a $n \times n$ positive definite covariance matrix Σ, if its probability density function is given by

$$f(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

   where $\mathrm{E}(X_i) = \mu_i$ and $cov(X_i, X_j) = \Sigma_{ij}$ for all $i, j = 1, \ldots, n$. We write this as $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

   In this question, we generalize the decision rule we have seen in the recitation to more than one dimension. Assume that the data is sampled from a known distribution $D(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{R}^d$ is a data point and y is its label. Denote the distributions $D(\mathbf{x}, 0)$ and $D(\mathbf{x}, 1)$ by $f_0(\mathbf{x})$ and $f_1(\mathbf{x})$, respectively. It is known that f_0, f_1 are multivariate Gaussian:

$$f_0(\mathbf{x}) = f(x; \boldsymbol{\mu}_0, \Sigma)$$
$$f_1(\mathbf{x}) = f(x; \boldsymbol{\mu}_1, \Sigma)$$

   Note that the covariance matrix, Σ, is the same for both distributions, but the mean vectors, $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1$, are different. Finally, it is known that the probability to sample a positive sample (i.e. $y = 1$) is p.

(a) We are given a point $\mathbf{x}$ and we need to label it with either $y = 0$ or $y = 1$. Suppose our decision rule is to decide $y = 1$ if and only if $\Pr(y = 1|\mathbf{x}) > \Pr(y = 0|\mathbf{x})$. Find a simpler condition for $\mathbf{x}$ that is equivalent to this rule.

(b) The decision boundary for this problem is defined as the set of points for which $\Pr(y = 1|\mathbf{x}) = \Pr(y = 0|\mathbf{x})$. What is the shape of the decision boundary when $d = 1$? When $d = 2$? For a general $d > 1$?

2. In this question, we will show that the Markov and Chebyshev inequalities are *tight* – that is, that they cannot be improved in general, given only the assumptions on which they rely.

(a) For any $a > 0$, provide a random variable $X \geq 0$ for which Markov's inequality is met with equality, i.e.:

$$\Pr(X \geq a) = \frac{\mathrm{E}(X)}{a}$$

(b) For any $b > 0$, provide a random variable X for which Chebyshev's inequality is met with equality, i.e.:

$$\Pr(|X - \mathrm{E}(X)| \geq b) = \frac{\mathrm{Var}(X)}{b^2}$$

3. Suppose we need to build a load balancing device to assign a set of n jobs to a set of m servers. Suppose the j-th job takes L_j time, $0 \leq L_j \leq 1$ (say, in seconds). The goal is to assign the n jobs to the m servers so that the load is as balanced as possible (i.e., so that the busiest server finishes as quickly as possible). Suppose each server works sequentially through the jobs that are assigned to it and finishes in time equal to the sum of job lengths assigned to the server. Let $L = \sum_{j=1}^{n} L_j$ be the total sum of job lengths (assume $L >> m$). With perfect load balancing, each server would take L/m time. There are some good algorithms for this scenario, but we are interested in analyzing the case of random assignment of jobs to servers.

Suppose we assign a random server for each job, with replacement. Denote by $R_{i,j}$ the load on server i from job j - that is, L_j if server i was assigned for job j, or 0 otherwise. Also, let $R_i = \sum_{j=1}^{n} R_{ij}$ be the total load on server i.

(a) What is $\mathrm{E}(R_i)$?

(b) We want to bound the probability that the load on the i-th server is more than $\delta = 10\%$ larger than the expected load. Use the multiplicative form of the Chernoff bound (see recitation scribe) to bound

$$\Pr(R_i \geq (1 + \delta) \cdot \mathrm{E}(R_i))$$

Note that this form doesn't require the summed random variables to be identically distributed.

(c) Now, we want to bound the probability that **any** of the servers are overloaded by more than $\delta = 10\%$ of the expected load. Give a bound for:

$$\Pr\left(R_1 \geq (1 + \delta) \cdot \mathrm{E}(R_1) \;\; or \;\; \ldots \;\; or \;\; R_m \geq (1 + \delta) \cdot \mathrm{E}(R_m)\right)$$

using the results from (a) and (b) and using the union bound (reminder: for events $A_1, \ldots, A_k$, the union bound is $\Pr(A_1 \cup \ldots \cup A_k) \leq \sum_{i=1}^{k} \Pr(A_i)$).