CSC412/2056 Assignment #1

Problem 1 (Variance and covariance, 6 points)
Let X and Y be two continuous independent random variables.

(a) Starting from the definition of independence, show that the independence of X and Y implies that their covariance is zero.

(b) For a scalar constant a, show the following two properties, starting from the definition of expectation:

$$\mathbb{E}(X + aY) = \mathbb{E}(X) + a\mathbb{E}(Y)$$
$$\text{var}(X + aY) = \text{var}(X) + a^2\text{var}(Y)$$

Problem 2 (Densities, 5 points)
Answer the following questions:

(a) Can a probability density function (pdf) ever take values greater than 1?

(b) Let X be a univariate normally distributed random variable with mean 0 and variance $1/100$. What is the pdf of X?

(c) What is the value of this pdf at 0?

(d) What is the probability that $X = 0$?

(e) Explain the discrepancy.

Problem 3 (Calculus, 4 points)
Let $x, y \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times m}$. Please answer the following questions, writing your answers in vector notation.

(a) What is the gradient with respect to x of $x^T y$?

(b) What is the gradient with respect to x of $x^T x$?

(c) What is the gradient with respect to x of $x^T A$?

(d) What is the gradient with respect to x of $x^T A x$?

Problem 4 (Linear Regression, 10pts)

Suppose that $X \in \mathbb{R}^{n\times m}$ with $n \geq m$ and $Y \in \mathbb{R}^n$, and that $Y \sim \mathcal{N}(X\beta, \sigma^2 I)$. In this question you will derive the result from class that the maximum likelihood estimate $\hat{\beta}$ of β is given by

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

(a) Why do we need to assume that $n \geq m$?

(b) What are the expectation and covariance matrix of $\hat{\beta}$, for a given true value of β?

(c) Show that maximizing the likelihood is equivalent to minimizing the squared error $\sum_{i=1}^n (y_i - x_i\beta)^2$. [Hint: Use $\sum_{i=1}^n a_i^2 = a^T a$]

(d) Write the squared error in vector notation, (see above hint), expand the expression, and collect like terms. [Hint: Use $\beta^T x^T y = y^T x\beta$ (why?) and $x^T x$ is symmetric]

(e) Take the derivative of this expanded expression with respect to β to show the maximum likelihood estimate $\hat{\beta}$ as above. [Hint: Use results 3.c and 3.d for derivatives in vector notation.]

Problem 5 (Ridge Regression, 5pts)

Suppose we place a normal prior on β. That is, we assume that $\beta \sim \mathcal{N}(0, \tau^2 I)$. The MAP estimate of β given Y in this context is

$$\hat{\beta}_{MAP} = (X^T X + \lambda I)^{-1} X^T Y$$

where $\lambda = \sigma^2/\tau^2$.

Estimating β in this way is called *ridge regression* because the matrix λI looks like a "ridge". Ridge regression is a common form of *regularization* that is used to avoid the overfitting that happens when the sample size is close to the output dimension in linear regression.

(a) Do we need $n \geq m$ to do ridge regression? Why or why not?

(b) Show that ridge regression is equivalent to adding m additional rows to X where the j-th additional row has its j-th entry equal to $\sqrt{\lambda}$ and all other entries equal to zero, adding m corresponding additional entries to Y that are all 0, and and then computing the maximum likelihood estimate of β using the modified X and Y.

Problem 6 (High dimensions, 5pts)

A hypersphere is the generalization of the concept of a sphere, to arbitrary dimension (not just $d = 3$). Consider a d-dimensional hypersphere of radius r. The fraction of its hypervolume lying between values $r - c$ and r, where $0 < c < r$, is given by

$$f = 1 - \left(1 - \frac{c}{r}\right)^d .$$

(a) For any fixed c value, f tends to 1 as $d \to \infty$. Show this numerically, with $c/r = 0.01$, for $d = 2$, 10, and 1000.

(b) Evaluate the fraction of the hypervolume which lies inside the radius $r/2$ for $d = 2$, 10, and 1000.

(c) For uniformly distributed points inside a high-dimensional hypersphere centred at the origin, select one of the following as correct (no explanation needed):

   (a) Most points are found near the middle of the hypersphere (the origin),

   (b) Most points are found along the axes,

   (c) Most points are found close to the hypersphere's surface, or

   (d) None of the above.