Phileas Hocquard
phileas@cs.toronto.edu
CSC412/2056

Probabilistic Learning and Reasoning, #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.**

As X is continuous, the mean $\mathbb{E}[X] = \int_x x * p(x)dx$

Where the variance of $\text{var}(X) = \mathbb{E}[(X - [E[X])^2]$ = $\int_x (x - \mathbb{E}[x])^2 \cdot p(x)dx = \mathbb{E}[x^2] - \mathbb{E}[x]^2$

$\text{cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$ = $\mathbb{E}(XY) - \mathbb{E}[X]\mathbb{E}[Y]$
as $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$

Therefore,

$$\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$
$$\text{cov}(X, Y) = \int_{\mathbb{R}}\int_{\mathbb{R}} F_{(X,Y)}(x, y) - F_X(x)F_Y(y)dxdy$$
$$\text{cov}(X, Y) = \int_{\mathbb{R}}(\int_{\mathbb{R}} f(x)dx)f(y)dy - \int_{\mathbb{R}} f(x)dx \cdot \int_{\mathbb{R}} f(y)dy$$
$$Since \int_{\mathbb{R}}(\int_{\mathbb{R}} f(x)dx)f(y)dy = \int_{\mathbb{R}} f(x)dx \cdot \int_{\mathbb{R}} f(y)dy$$
$$\text{cov}(X, Y) = 0$$

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

(a) $\mathbb{E}[X + aY] = E(X) + a\mathbb{E}(Y)$

$\mathbb{E}[X + aY] = \int_{\mathbb{R}}\int_{\mathbb{R}}(x + ay)f_{(x,y)}(x, y)dxdy$

$\mathbb{E}[X + aY] = \int_{\mathbb{R}}\int_{\mathbb{R}}(x)f_{(x,y)}(x, y)dxdy$ + $\int_{\mathbb{R}}\int_{\mathbb{R}} ayf_{(x,y)}(x, y)dxdy$

$\mathbb{E}[X + aY] = \int_{\mathbb{R}} x \int_{\mathbb{R}} f_{(x,y)}(x, y)dydx$ + $a \int_{\mathbb{R}} y \int_{\mathbb{R}} f_{(x,y)}(x, y)dxdy$

Through marginalization of x and of y,
we can simplify $\int_{\mathbb{R}} f_{(x,y)}(x, y)dy$, $\int_{\mathbb{R}} f_{(x,y)}(x, y)dx$.

$\mathbb{E}[X + aY] = \int_{\mathbb{R}} xf_x(x)dx$ + $a \int_{\mathbb{R}} yf_y(y)dy$

$\mathbb{E}[X + aY] = E(X) + a\mathbb{E}(Y)$

(b) $\textbf{var}(X + aY) = \textbf{var}(X) + a^2\textbf{var}(Y)$

i. Firstly,
$\mathbb{E}[(X + aY)^2] = \int_{\mathbb{R}} \int_{\mathbb{R}} (x + ay)^2 f_{(x,y)}(x, y)dxdy$
$\mathbb{E}[(X + aY)^2] = \int_{\mathbb{R}} \int_{\mathbb{R}} (x)^2 + (ay)^2 + 2(xay) f_{(x,y)}(x, y)dxdy$
= $\mathbb{E}[X^2] + 2a\mathbb{E}[XY] + a^2\mathbb{E}[Y^2]$
= $\mathbb{E}[X^2] + 2a\mathbb{E}[X]\mathbb{E}[Y] + a^2\mathbb{E}[Y^2]$

ii. Secondly,
$\mathbb{E}[(X + aY)]^2 = (\mathbb{E}[X] + a\mathbb{E}[Y])^2$
= $\mathbb{E}[X]^2 + 2a\mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[Y]^2$

iii. As $\text{var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2$
$\text{var}(X + aY) = \mathbb{E}[(X + aY)^2] - \mathbb{E}[(X + aY)]^2$
$\text{var}(X + aY) = \mathbb{E}[X^2] + 2a\mathbb{E}[X]\mathbb{E}[Y] + a^2\mathbb{E}[Y^2] - (\mathbb{E}[X]^2 + 2a\mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[Y]^2)$
As we can see the terms $2a\mathbb{E}[X]\mathbb{E}[Y]$ and $2a\mathbb{E}[Y]\mathbb{E}[X]$ cancel out.

iv. Therefore we are left with:
$\text{var}(X + aY) = (\mathbb{E}[X^2] - \mathbb{E}[X]^2) + a^2(\mathbb{E}[Y^2] - \mathbb{E}[Y]^2)$
$\text{var}(X + aY) = \text{var}(X) + a^2\text{var}(Y)$

Problem 2 (Densities, 5 points)

(a) **Can a probability density function (pdf) ever take values greater than 1?**
Yes, as probability density function describes a probability density and not a probability therefore the value can be greater than one.

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

$pdf(X) = f(x|\mu, \sigma^2) = \frac{1}{\sqrt{\frac{\pi}{50}}} e^{-\frac{(x)^2}{\frac{1}{50}}}$

(c) **What is the value of this pdf at 0?**
The value of the pdf for a univariate normal distribution at zero for the given values is $\frac{1}{\sqrt{\frac{\pi}{50}}}$

(d) **What is the probability that $X = 0$?**
The given value is Zero.
$\int_0^0 \frac{1}{\sqrt{\frac{\pi}{50}}} e^{-\frac{(x)^2}{\frac{1}{50}}} = 0$. As for any integration using the same specific value for the upper and lower bound we will always obtain 0 since there is no area under the curve.

(e) **Explain the discrepancy.**
The difference in results here is simply that one determines the density which corresponds to probability per unit value of random variable and the other tries to determine a single probability. As we can see we obtain approximately 3.98 as a pdf and 0 as a probability.

Problem 3 (Calculus, 4 points)
Let $x, y \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times m}$.

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

$$\nabla_x f = \frac{\partial f}{\partial x} = \frac{\partial (x^T y)}{\partial x} = \frac{\partial [x_1 x_2 \cdots x_m] \begin{bmatrix} y_1 \\ y_2 \\ y_m \end{bmatrix}}{\partial x} = \frac{\partial s}{\partial x}$$

$$\nabla_x f = \frac{\partial s}{\partial x} = \begin{bmatrix} \frac{\partial s}{\partial x_1} \\ \frac{\partial s}{\partial x_2} \\ \vdots \\ \frac{\partial s}{\partial x_m} \end{bmatrix} = [y_1 y_2 \cdots y_m] = (y_1 y_2 \cdots y_m)^T, as \frac{\partial s}{x_1} = y_1, \frac{\partial s}{x_2} = y_2, \cdots, \frac{\partial s}{x_m} = y_m$$

$$Therefore, \nabla_x (x^T y) = y^T$$

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

$$\nabla_x f = \frac{\partial f}{\partial x} = \frac{\partial (x^T x)}{\partial x} = \frac{\partial [x_1 x_2 \cdots x_m] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix}}{\partial x} \rightarrow \frac{\partial \sum_{i=1}^{n} x_i^2}{\partial x_i} = 2\mathrm{x}_i$$

$$\rightarrow \frac{\partial (x^T x)}{\partial x} = (2\mathrm{x}_1, 2\mathrm{x}_2, \cdots, 2\mathrm{x}_m) = 2\mathrm{x}^T$$

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

$$x=\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}, A=\begin{pmatrix} a_{11}\, a_{12} \cdots a_{1\mathrm{m}} \\ a_{21} \ddots \cdots a_{2\mathrm{m}} \\ \vdots \ldots \ddots \ddots \ddots \ldots \vdots \\ a_{1\mathrm{m}} \ldots\ldots \ddots\, a_{nm} \end{pmatrix}, \quad s=x^T A=(x_1\, x_2 \cdots x_m)\begin{pmatrix} a_{11}\, a_{12} \cdots a_{1\mathrm{m}} \\ a_{21} \ddots \cdots a_{2\mathrm{m}} \\ \vdots \ldots \ddots \ddots \ddots \ldots \vdots \\ a_{1\mathrm{m}} \ldots\ldots \ddots\, a_{nm} \end{pmatrix}$$

$$s=\begin{pmatrix} a_{11} x_1 + a_{12} x_2 \cdots + a_{1\mathrm{m}} x_m \\ a_{21} x_1 + \cdots \ddots \cdots + a_{2\mathrm{m}} x_m \\ \vdots \ldots\ldots\ldots\ldots\ldots \ddots \ddots \ddots \ldots \vdots \\ a_{1\mathrm{m}} x_1 + \cdots\cdots\cdots \ddots x_{nm} x_m \end{pmatrix} = \sum_{j=1}^{m} \sum_{i=1}^{n} a_{ij} x_i$$

$$\nabla_x f = \frac{\partial x^T A}{\partial x} = \frac{\partial s}{\partial x} = A^T, as \frac{\partial s}{x_i} = \sum_{j=1}^{m} \sum_{i=1}^{n} a_{ij}, where\ for\ any\ given\ ij \frac{\partial a_{ij} x_i}{x_i} = a_{ij}$$

$$Therefore, \frac{\partial x^T A}{\partial x} = A^T$$

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

$$x=\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}, A=\begin{pmatrix} a_{11}\, a_{12} \cdots a_{1\mathrm{m}} \\ a_{21} \ddots \cdots a_{2\mathrm{m}} \\ \vdots \ldots \ddots \ddots \ddots \ldots \vdots \\ a_{1\mathrm{m}} \ldots\ldots \ddots\, a_{nm} \end{pmatrix}$$

$$s=x^T Ax=(x_1\, x_2 \cdots x_m)\begin{pmatrix} a_{11}\, a_{12} \cdots a_{1\mathrm{m}} \\ a_{21} \ddots \cdots a_{2\mathrm{m}} \\ \vdots \ldots \ddots \ddots \ddots \ldots \vdots \\ a_{1\mathrm{m}} \ldots\ldots \ddots\, a_{nm} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

$$s=(x_1\, x_2 \cdots x_m)\begin{pmatrix} a_{11} x_1 + a_{12} x_2 \cdots + a_{1\mathrm{m}} x_m \\ a_{21} x_1 + \cdots \ddots \cdots + a_{2\mathrm{m}} x_m \\ \vdots \ldots\ldots\ldots\ldots\ldots \ddots \ddots \ddots \ldots \vdots \\ a_{1\mathrm{m}} x_1 + \cdots\cdots\cdots \ddots a_{nm} x_m \end{pmatrix} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

$$\nabla_x f = \frac{\partial x^T Ax}{\partial x} = \frac{\partial s}{\partial x} = x^T A^T + x^T A, as \frac{\partial s}{x_k} = \sum_{j=1}^{n} a_{kj} x_j + \sum_{i=1}^{n} a_{ki} x_i$$

$$Therefore, \frac{\partial x^T Ax}{\partial x} = x^T (A^T + A)$$

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 $n \geq m$?**

Imagining that we were to compute a left inverse on each side of the $\hat{\beta}$ matrix.
If we look at $\hat{\beta}$ as a system of **n** equations with **m** unknowns. There is no unique solution to this system when the number of unknowns **m** is larger than **n**. This can lead to over-fitting.

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

(a) Expectation of $\hat{\beta}$ given β

$$\begin{aligned}
\mathbb{E}[\hat{\beta}] &= \mathbb{E}[(X^T X)^{-1} X^T Y], Y = X\beta + \epsilon \\
&= \mathbb{E}[(X^T X)^{-1} X^T (X\beta + \epsilon)] \\
&= \mathbb{E}[(X^T X)^{-1} X^T] + \mathbb{E}[(X\beta + \epsilon)] \\
&= \mathbb{E}[(X^T X)^{-1} X^T X\beta + ((X^T X)^{-1} X^T \epsilon)] \\
&= \mathbb{E}[I\beta](+\mathbb{E}[\epsilon]), were \mathbb{E}[\epsilon] = 0 \\
&= \mathbb{E}[\beta] \\
\mathbb{E}[\hat{\beta}] &= \mathbb{E}[\beta]
\end{aligned}$$

(b) Covariance of $\hat{\beta}$ given β

$$\begin{aligned}
\text{cov}[\hat{\beta}] &= \text{cov}[(X^T X)^{-1} X^T (X\beta + \epsilon)] \\
&= \text{cov}[\beta + (X^T X)^{-1} x^T \epsilon],\ \text{var}[\beta] = \text{cov}[\beta] = 0 \\
&= \cancel{\text{cov}[\beta]} + \text{cov}[(X^T X)^{-1} X^T \epsilon] \\
&= (X^T X)^{-1} X^T \text{cov}[\epsilon] X (X^T X)^{-1} \\
&= \text{cov}[\epsilon] \cancel{(X^T X)^{-1} (X^T X)} (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1}
\end{aligned}$$

.

.

(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 $\sum_{i=1}^{n} a_i^2 = a^T a$]**

$$\textit{The likelihood is given by } \prod_{i=1}^{n} \frac{e^{-\frac{(y_i - x_i\beta)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} = \prod_{i=1}^{n} f(x_i|\beta)$$

By maximizing the log *likelihood we obtain the same as minimizing the negative* log *likelihood.*
The log *likelihood* :

$$\zeta = \log \textit{likelihood} = \sum_{i=1}^{m} \log f(x_i|\beta)$$

$$max_\beta(\zeta) = min_\beta(-\zeta)$$

$$max_\beta(\zeta) = min_\beta(-\sum_{i=1}^{m} \log f(x_i|\beta))$$

$$max_\beta(\zeta) = min_\beta(\sum_{i=1}^{m} \frac{-(y_i - x_i\beta)^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma))$$

$$max_\beta(\zeta) = min_\beta(n\log(\sqrt{2\pi}\sigma) - \sum_{i=1}^{m} \frac{-1}{2\sigma^2}(y_i - x_i\beta)^2)$$

$$max_\beta(\zeta) = n\log(\sqrt{2\pi}\sigma) + \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - x_i\beta)^2$$

As we can see on the right side of the equation we obtain the minimizing squared error function and the left side a constant.

(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]**

$$\sum_{i=1}^{m}(y_i - x_i\beta)^2 = \sum_{i=1}^{m}(y_i - x_i\beta)(y_i - x_i\beta)$$

$$\sum_{i=1}^{m}(y_i - x_i\beta)^2 = \sum_{i=1}^{m}(y_i^2 + (x_i\beta)^2 - 2y_i x_i\beta)$$

$$\textit{Since } \beta^T x^T y = y^T x\beta$$

$$\sum_{i=1}^{m}(y_i - x_i\beta)^2 = y^T y + x^T\beta^T\beta x - 2\beta^T x^T y$$

The importance behind this symmetric property is in regards to when the matrix dimensionality has to be conserved when deriving.

(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.]**

$$\frac{\partial\tilde{\beta}}{\partial\beta}=\frac{\partial Y^T Y+X^T\beta^T\beta X-2\beta^T X^T Y}{\partial\beta}=2\mathrm{X}^T X\beta-2\mathrm{X}^T Y$$
$$\frac{\partial\tilde{\beta}}{\partial\beta}=0=2\mathrm{X}^T X\beta-2\mathrm{X}^T Y$$
$$\frac{\partial\tilde{\beta}}{\partial\beta}=(2\mathrm{X}^T X\beta=2\mathrm{X}^T Y)$$
$$\frac{\partial\tilde{\beta}}{\partial\beta}=(X^T X\beta=X^T Y)$$
$$\beta'=(X^T X)^{-1}X^T Y$$

Problem 5 (Ridge Regression, 5pts)

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

No.
As previously mentioned (for the linear regression question). If we were to compute as an expression of system of equations for an inverse matrix, we need to establish a balance between the **m** and **n** dimensions. Fortunately, the terms λ I does exactly this. Regularization solves what is called an ill-posed problem. For a system of equation is unbalanced for a matrix **A**, such as **Ax = b** satisfies no values or fit more than one property, adding terms λ **I** as the following $\|Ax\| + \|\lambda I\|$ will allow a balance on the system of equation.

(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.**

$$Y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1z} \\ \vdots & \ddots & x_{nz-1} & x_{nz} \end{pmatrix}, \sqrt{\lambda} I=\begin{pmatrix} \sqrt{\lambda} & 0 & \cdots & 0_{1g} \\ 0 & \sqrt{\lambda} & 0 & \vdots \\ 0 & \ddots & \ddots & \vdots \\ 0_{m1} & 0 & \cdots & \sqrt{\lambda_{mg}} \end{pmatrix}$$
$$Y'=\begin{bmatrix} Y \\ 0_m \end{bmatrix}, X'=\begin{bmatrix} \underset{n}{X} \\ \underset{m}{\sqrt{\lambda} I} \end{bmatrix}$$
$$From.(X^T X)^{-1}X^T Y$$
$$X'^T X'=[X^T\sqrt{\lambda} I]\begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix}=(X^T X+\lambda I)$$
$$Therefore, we\ can\ obtain (X^T X+\lambda I)^{-1}X'^T Y'$$