Homework Set 1

Introduction to Artificial Intelligence with Mathematics (MAS473)

1. (10pts) The ABC test is a new test for diagnosing depression. An extensive clinical evaluation was performed of this instrument, whereby participants were interviewed by psychiatrists and a definitive clinical diagnosis of depression was made. The table below shows the number of participants with or without depression based on test scores. If one gets the test score higher than x, we are going to diagnose depression for him/her.

	With depression (positive)	Without depression (negative)
$0 \sim 5$	0	6
$6 \sim 10$	1	20
$11 \sim 15$	1	9
$16 \sim 20$	3	4
$21 \sim 25$	5	1

   (a) If $x = 15$, find the sensitivity and the specificity.

   (b) Calculate F1 score for each $x = 5, 10, 15, 20$ and find the best value x among $5, 10, 15, 20$ with respect to the F1 score.

2. (10pts) A distribution $\mathrm{Dir}(\alpha)$ (called a Dirichlet distribution) is a distribution whose probability density function is given by

$$\mathrm{Dir}(\mathbf{x}|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mathbf{x}_k^{\alpha_k - 1}$$

on $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_K)^\top \in \triangle^{K-1} = \{(y_1, \cdots, y_K) \in \mathbb{R}^K : y_1 + \cdots + y_K = 1 \text{ and } y_i \geq 0 \text{ for any } i = 1, \cdots, K\}$ where $\alpha = (\alpha_1, \cdots, \alpha_K)$ ($\alpha_i > 0$ for any i), $\alpha_0 = \alpha_1 + \cdots + \alpha_K$, and the gamma function Γ is defined by

$$\Gamma(y) = \int_0^\infty t^{y-1} e^{-t} \, dt.$$

   (a) Prove that $\mathrm{Dir}(\alpha)$ is normalized, i.e.

$$\int_{\triangle^{K-1}} \mathrm{Dir}(\mathbf{x}|\alpha) \, d\mathbf{x} = 1.$$

   You can use the fact that the Beta distribution is normalized.

   (b) Find $\mathbf{E}[\mathbf{x}_i]$, $\mathrm{Var}(\mathbf{x}_j)$ and $\mathrm{Cov}(\mathbf{x}_i, \mathbf{x}_j)$ ($i \neq j$) where $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_K) \sim \mathrm{Dir}(\alpha)$.
   (Hint : You may use a property of gamma function $\Gamma(x+1) = x\Gamma(x)$ for $x > 0$)

3. (10pts) A K-dimensional multivariate normal distribution $\mathcal{N}(\mu, \mathbf{\Sigma})$ is a distribution whose probability density function is given by

$$\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{K/2}} \cdot \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)\right)$$

on $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_K) \in \mathbb{R}^K$ where $\mu = (\mu_1, \cdots, \mu_K)^\top \in \mathbb{R}^K$ and $\Sigma \in \mathbb{R}^{K\times K}$ is a positive definite matrix. Let $X_1, \cdots, X_n$ be i.i.d. random variables with the probability density function $\mathcal{N}(\mu, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is known.

(a) Find the maximum likelihood estimator (MLE) of μ.

(b) Suppose μ has a prior distribution $p(\mu) = \mathcal{N}(\mu|\mu_0, \mathbf{\Sigma}_0)$. Find the posterior distribution of μ and the maximum a posterior (MAP) estimator of μ.

4. (10pts) Answer the followings.

(a) Let X be a discrete random variable that has a probability mass

$$\mathbf{P}(X = x_i) = p_i \quad (i = 1, \cdots, m)$$

for $p_i > 0$ and $\sum_{i=1}^m p_i = 1$. Prove that

$$H(X) \le \log m$$

where $H(X)$ denotes the entropy of X.

(b) Calculate the KL divergence between two univariate Gaussian distributions $D_{KL}(\mathcal{N}(\mu, \sigma^2) \parallel \mathcal{N}(m, s^2))$.

5. (10pts) Note that there are two different layout conventions : the numerator-layout notation and the denominator-layout notation. In our lecture, we use the numerator-layout notation. With the numerator-layout notation,

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial \mathbf{x}_1} & \frac{\partial y}{\partial \mathbf{x}_2} & \cdots & \frac{\partial y}{\partial \mathbf{x}_n} \end{bmatrix}, \qquad \frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial \mathbf{y}_1}{\partial x} & \frac{\partial \mathbf{y}_2}{\partial x} & \cdots & \frac{\partial \mathbf{y}_m}{\partial x} \end{bmatrix}^\top$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_2} & \cdots & \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_n} \\ \frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_2} & \cdots & \frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_2} & \cdots & \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_n} \end{bmatrix}, \qquad \frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial \mathbf{X}_{11}} & \frac{\partial y}{\partial \mathbf{X}_{21}} & \cdots & \frac{\partial y}{\partial \mathbf{X}_{p1}} \\ \frac{\partial y}{\partial \mathbf{X}_{12}} & \frac{\partial y}{\partial \mathbf{X}_{22}} & \cdots & \frac{\partial y}{\partial \mathbf{X}_{p2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial \mathbf{X}_{1q}} & \frac{\partial y}{\partial \mathbf{X}_{2q}} & \cdots & \frac{\partial y}{\partial \mathbf{X}_{pq}} \end{bmatrix},$$

$$\frac{\partial \mathbf{Y}}{\partial x} = \begin{bmatrix} \frac{\partial \mathbf{Y}_{11}}{\partial x} & \frac{\partial \mathbf{Y}_{12}}{\partial x} & \cdots & \frac{\partial \mathbf{Y}_{1n}}{\partial x} \\ \frac{\partial \mathbf{Y}_{21}}{\partial x} & \frac{\partial \mathbf{Y}_{22}}{\partial x} & \cdots & \frac{\partial \mathbf{Y}_{2n}}{\partial x} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{Y}_{m1}}{\partial x} & \frac{\partial \mathbf{Y}_{m2}}{\partial x} & \cdots & \frac{\partial \mathbf{Y}_{mn}}{\partial x} \end{bmatrix}$$

where

$$x, y \in \mathbb{R},\ \mathbf{x} \in \mathbb{R}^n,\ \mathbf{y} \in \mathbb{R}^m,\ \mathbf{X} \in \mathbb{R}^{p\times q},\ \mathbf{Y} \in \mathbb{R}^{m\times n}.$$

On the other hand, with the denominator-layout notation,

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial \mathbf{x}_1} & \frac{\partial y}{\partial \mathbf{x}_2} & \cdots & \frac{\partial y}{\partial \mathbf{x}_n} \end{bmatrix}^\top, \qquad \frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial \mathbf{y}_1}{\partial x} & \frac{\partial \mathbf{y}_2}{\partial x} & \cdots & \frac{\partial \mathbf{y}_m}{\partial x} \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_1} & \cdots & \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_1} \\ \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_2} & \frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_2} & \cdots & \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{y}_1}{\partial \mathbf{x}_n} & \frac{\partial \mathbf{y}_2}{\partial \mathbf{x}_n} & \cdots & \frac{\partial \mathbf{y}_m}{\partial \mathbf{x}_n} \end{bmatrix}, \qquad \frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial \mathbf{X}_{11}} & \frac{\partial y}{\partial \mathbf{X}_{12}} & \cdots & \frac{\partial y}{\partial \mathbf{X}_{1q}} \\ \frac{\partial y}{\partial \mathbf{X}_{21}} & \frac{\partial y}{\partial \mathbf{X}_{22}} & \cdots & \frac{\partial y}{\partial \mathbf{X}_{2q}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial \mathbf{X}_{p1}} & \frac{\partial y}{\partial \mathbf{X}_{p2}} & \cdots & \frac{\partial y}{\partial \mathbf{X}_{pq}} \end{bmatrix}$$

where

$$x, y \in \mathbb{R},\ \mathbf{x} \in \mathbb{R}^n,\ \mathbf{y} \in \mathbb{R}^m,\ \mathbf{X} \in \mathbb{R}^{p\times q},\ \mathbf{Y} \in \mathbb{R}^{m\times n}.$$

(Please read `https://en.wikipedia.org/wiki/Matrix_calculus#Layout_conventions` for details.) In this problem, we use the numerator-layout notation as in the lecture. Answer the followings.

(a) Prove that

$$\frac{\partial}{\partial A}\mathrm{tr}(AB) = B$$

for $n \times n$ matrices A and B. In particular,

$$\frac{\partial}{\partial A}\mathrm{tr}(A) = I.$$

(b) Prove that

$$\frac{\partial}{\partial x}\log|A| = \mathrm{tr}\left(A^{-1}\frac{\partial A}{\partial x}\right)$$

for a $n \times n$ invertible matrix $A(x)$ where $x \in \mathbb{R}$ and $|A| = \det(A) > 0$.
(Hint: Prove when A is symmetric, and then prove the general case. Note that if A is symmetric, we can represent

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^\top$$

where λ_i is an eigenvalue of A, u_i is a corresponding eigenvector, and u_i's are orthonormal.)

(c) Prove that

$$\frac{\partial}{\partial A}\log|A| = A^{-1}$$

for a $n \times n$ invertible matrix A where $|A| = \det(A) > 0$.