EECS 545 – Machine Learning - Homework #1

David Ke Hong — Due: 11:00pm 01/25/2016

1) **Linear Algebra (25 pts).**

(a) (15 pts)

(i) True **(2 pts)**, $I = (A^{-1}A)^\top = A^\top(A^\top)^{-1}$, thus $A^\top(A^{-1})^\top = A^\top(A^\top)^{-1}$ and $(A^{-1})^\top = (A^\top)^{-1}$. **(3 pts)**

(ii) False **(2 pts)**, let $A = B = I$, $A^{-1} = B^{-1} = I$ and $(A+B)^{-1} \neq 2I$. **(3 pts)**

(iii) True **(2 pts)**, given a symmetric, invertible matrix A, $A = A^\top$ and thus $A^{-1} = (A^\top)^{-1} = (A^{-1})^\top$, according to part (i). **(3 pts)**

(b) (5 pts) $\Lambda = \Sigma\Sigma^\top$, which is a diagonal $m \times m$ matrix with the squares of the singular values of X along the diagonal. **(2.5 pts)**
$Q = U$. **(2.5 pts)**

(c) (5 pts)

(i) 757.00, 158.21, 130.32 **(3 pts)**

(ii) 68125 ± 682 **(2 pts)**

2) **Probability (20 pts).**

(a) (8 pts)

(i) Depends on the overlap of the distributions of D and H. **(2 pts)**
Given a very small overlap, the relation will be "$\leq$". If the event $D = d$ is a subset of the event $H = h$ in the sample space, however, the relation will be "$\geq$". **(2 pts)**

(ii) The relation is "$\geq$". **(2 pts)**
The left-hand side is equal to the right hand side divided by $P(D = d)$ by Bayes' rule, and since $P(D = d) \leq 1$, the relation is "$\geq$". **(2 pts)**

(b) (12 pts)

(i) $\mathbb{E}[X] = \int\int xp(x,y)\mathrm{d}x\mathrm{d}y = \int(\int xp(x|y)\mathrm{d}x)p(y)\mathrm{d}y = \mathbb{E}_Y[\mathbb{E}_X[X|Y]]$ **(4 pts)**

(ii) $\mathrm{var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$
$= \mathbb{E}[X^2] - \mathbb{E}[X]^2$
$= \mathbb{E}_Y[\mathbb{E}_X[X^2|Y]] - \mathbb{E}[X]^2$
$= \mathbb{E}_Y[\mathbb{E}_X[X^2|Y]] - (\int xp(x|y)\mathrm{d}x)^2 + (\int xp(x|y)\mathrm{d}x)^2) - \mathbb{E}[X]^2$
$= \mathbb{E}_Y[\mathbb{E}_X[X^2|Y]] - \mathbb{E}_X[X|Y]]^2] + \mathbb{E}_Y[\mathbb{E}_X[X|Y]]^2] - (\mathbb{E}_Y[\mathbb{E}_X[X|Y]])^2$
$= \mathbb{E}_Y[\mathrm{var}_X[X|Y]] + \mathrm{var}_Y[\mathbb{E}_X[X|Y]]$ **(8 pts)**

3) **Positive (Semi-)Definite Matrices (20 pts).**

(a) First suppose A is positive semi-definite, then we have each $i = 1, ..., d$,

$$\lambda_i = \lambda_i \mathbf{u}_i^\top \mathbf{u}_i = \mathbf{u}^\top(\lambda_i \mathbf{u}_i) = \mathbf{u}^\top(A\mathbf{u}_i) \geq 0$$

where the last inequality follows from the positive semi-definite assumption. **(5pts)**
Now suppose $\lambda_i \geq 0$ for each i, then we have for all $\mathbf{x} \in \mathbb{R}^d$,

$$\mathbf{x}^\top A\mathbf{x} = \mathbf{x}^\top(U\Lambda U^\top)\mathbf{x} = \mathbf{x}^\top(\sum_{i=1}^{d} \lambda_i \mathbf{u}_i \mathbf{u}_i^\top)\mathbf{x} = \sum_{i=1}^{d} \lambda_i(\mathbf{x}^\top \mathbf{u})(\mathbf{u}_i^\top \mathbf{x}) = \sum_{i=1}^{d} \lambda_i \|\mathbf{u}_i^\top \mathbf{x}\|_2^2 \geq 0$$

Therefore, by definition A is positive semi-definite. **(5pts)**

(b) **(10pts)** The proof for the positive definite case is nearly identical to part (a). Note that to prove the converse implication, we must additionally assume that $\mathbf{x} \neq \mathbf{0}$.

4) **Maximum Likelihood Estimation (15 pts).** Recall the probability mass function of the Poisson distribution $Poi(\lambda)$ is given by:

$$f(\mathbf{X}; \lambda) = Pr(\mathbf{X} = x; \lambda) = \frac{\lambda e^{-\lambda}}{x!}$$

where $x \in 0, 1, 2, ...$ and $\lambda > 0$. Therefore, the log-likelihood function $\ell(\lambda)$ is

$$\ell(\lambda) = \sum_{i=1}^{n} \log f(\mathbf{X}_i; \lambda) = \sum_{i=1}^{n} \log(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}) = \sum_{i=1}^{n} x_i \log \lambda - n\lambda - \sum_{i=1}^{n} log x_i!$$

(6 pts)
Recognizing that maximizing $\ell(\lambda)$ is equivalent to minimizing its negative $-\ell(\lambda)$, the *maximum likelihood estimate* (MLE) of the parameter can be written as:

$$\hat{\lambda} \in \arg\min_{\lambda} -\ell(\lambda)$$

The first derivative of this objective function is

$$-\frac{\partial \ell \lambda}{\partial \lambda} = -\frac{\sum_{i=1}^{n} x_i}{\lambda} + n$$

whereas the second derivative is

$$-\frac{\partial^2 \ell \lambda}{\partial \lambda^2} = \frac{\sum_{i=1}^{n} x_i}{\lambda^2}$$

(6 pts)

Since this second derivative is nonnegative for all $\lambda > 0$, the negative log-likelihood is a convex function. Thus a global minimum of $-\ell(\lambda)$ can be found by setting its first derivative with respect to λ to zero, which give us the MLE:

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

(2 pts)

Notice if $x_i \neq 0$ for any $i \in 1, ..., n$, this MLE solution is unique due to the strict convexity of the objective function (the second derivative is strictly positive). On the other hand, an MLE does not exist if $x_i = 0$ for all i, as λ cannot be zero in a Poisson distribution ($\lambda > 0$). **(1 pts)**

5) **Unconstrained Optimization (20 pts).**

(a) **(5 pts)** We will prove this by contradiction. Suppose that f has two distinct global minimizers. By strict convexity, we have $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ where $\mathbf{x} \neq \mathbf{y}$ for any $t \in (0, 1)$

$$\begin{aligned} f(t\mathbf{x} + (1-t)\mathbf{y}) &< tf(\mathbf{x}) + (1-t)f(\mathbf{y}) \\ &= tf(\mathbf{x}) + (1-t)f(\mathbf{x}) \\ &= f(\mathbf{x}) \end{aligned}$$

This violates the global optimality of $\mathbf{x}$, establishing the contradiction.

(b) **(5 pts)** Let $\mathbf{y} \in \mathbb{R}^d$ be arbitrary, applying one of the quadratic expansions gives us for any $t \in \mathbb{R}$,

$$\begin{aligned} f(\mathbf{x}^* + t\mathbf{y}) &= f(\mathbf{x}^*) + \langle \bigtriangledown f(\mathbf{x}^*), t\mathbf{y}\rangle + \frac{t^2}{2}\langle \mathbf{y}, \bigtriangledown^2 f(\mathbf{x}^*)\mathbf{y}\rangle + o(t^2) \\ &= f(\mathbf{x}^*) + \frac{t^2}{2}\langle \mathbf{y}, \bigtriangledown^2 f(\mathbf{x}^*)\mathbf{y}\rangle + o(t^2) \end{aligned}$$

Rearranging, we have for suitciently small $t > 0$

$$0 \leq \frac{f(\mathbf{x}^* + t\mathbf{y}) - f(\mathbf{x}^*)}{t^2} = \frac{1}{2}\langle \mathbf{y}, \bigtriangledown^2 f(\mathbf{x}^*)\mathbf{y}\rangle + \frac{o(t^2)}{t^2}$$

where the inequality follows from the local minimality of $\mathbf{x}^*$. Finally, letting $t \to 0$ yields the inequality

$$0 \leq \langle \mathbf{y}, \bigtriangledown^2 f(\mathbf{x}^*)\mathbf{y}\rangle$$

Since $\mathbf{y} \in \mathbb{R}^d$ was arbitrary, this shows that $\bigtriangledown^2 f(\mathbf{x}^*)$ is positive semi-definite.

(c) **(5 pts)** First suppose f is convex, then for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $t \in \mathbb{R}$, we have

$$f(\mathbf{x} + t\mathbf{y}) \geq f(\mathbf{x}) + \langle \bigtriangledown f(\mathbf{x}), t\mathbf{y}\rangle$$

Applying one of the quadratic expansions to the left-hand side of this inequality gives

$$\begin{aligned} f(\mathbf{x} + t\mathbf{y}) &= f(\mathbf{x}) + \langle \bigtriangledown f(\mathbf{x}), t\mathbf{y}\rangle + \frac{t^2}{2}\langle \mathbf{y}, \bigtriangledown^2 f(\mathbf{x})\mathbf{y}\rangle + o(t^2) \\ &\geq f(\mathbf{x}) + \langle \bigtriangledown f(\mathbf{x}), t\mathbf{y}\rangle \end{aligned}$$

or equivalently,

$$\frac{1}{2}\langle \mathbf{y}, \bigtriangledown^2 f(\mathbf{x})\mathbf{y}\rangle + \frac{o(t^2)}{t^2} \geq 0$$

Since $\mathbf{x}$ and $\mathbf{y}$ were arbitrary, we have $\langle \mathbf{y}, \bigtriangledown^2 f(\mathbf{x})\mathbf{y}\rangle \geq 0$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, *i.e.*, $\bigtriangledown^2 f(\mathbf{x})$ is positive semi-definite for all $\mathbf{x} \in \mathbb{R}^d$. **(2.5 pts)**

Now suppose $\bigtriangledown^2 f(\mathbf{x})$ is positive semi-definite for all $\mathbf{x} \in \mathbb{R}^d$. Since f is twice continuously differentiable, we have for some $t \in (0, 1)$

$$f(\mathbf{x}) = f(\mathbf{y}) + \langle \bigtriangledown f(\mathbf{y})\mathbf{x} - \mathbf{y}\rangle + \frac{1}{2}\langle \mathbf{x} - \mathbf{y}, \bigtriangledown^2 f(\mathbf{y} + t(\mathbf{x} - \mathbf{y}))(\mathbf{x} - \mathbf{y})\rangle$$

Therefore, we have for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \bigtriangledown f(\mathbf{y}), (\mathbf{x} - \mathbf{y})\rangle$$

(2.5 pts)

(d) (5 pts) The quadratic function $f(\mathbf{x})$ can be written explicitly as:

$$\begin{aligned} f(x) &= \frac{1}{2}\mathbf{x}^\top A\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c \\ &= \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} A_{ij}x_i x_j + \sum_{i=1}^{d} b_i x_i + c \end{aligned}$$

Applying the definition of the Hessian matrix, the $(i, j)^{th}$ entry of $\bigtriangledown^2 f(\mathbf{x})$ is given by:

$$\begin{aligned} [\bigtriangledown^2 f(\mathbf{x})]_{i,j} &= \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \\ &= \frac{\partial^2}{\partial x_i \partial x_j}\{\frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} A_{ij}x_i x_j + \sum_{i=1}^{d} b_i x_i + c\} \\ &= \frac{1}{2}\frac{\partial^2}{\partial x_i \partial x_j}\{\frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} A_{ij}x_i x_j\} \\ &= A_{ij} \end{aligned}$$

thus the Hessian of f is A. **(3 pts)**
The function f is convex when A is positive semi-definite, and strictly convex if A is positive definite. **(2 pts)**