Homework 1 Solution

EECS 545 Machine Learning

September-22-2017

1) Linear Algebra

(a)

(i) True.

Proof. $I = A^{-1}A = (AA^{-1})^T = (A^{-1})^T A^T$. Because A is symmetric, $A^{-1}A = A^{-1}A^T = (A^{-1})^T A^T$. Right muliply by $(A^T)^{-1}$, we get $A^{-1} = (A^{-1})^T$. $\square$

(ii) True.

Proof. Assume an orthogonal matrix M has the form

$$M = \begin{bmatrix} p & q \\ r & t \end{bmatrix}$$

Due to the properties of orthogonal matrices, we have

$$\begin{cases} p^2 + q^2 = 1 \\ r^2 + t^2 = 1 \\ pr + qt = 0 \end{cases} \tag{1}$$

Without loss of generality, we can write $(p, q) = (\cos\theta, \sin\theta)$ or $(\cos\theta, -\sin\theta)$, and $(r, t) = (\cos\phi, \sin\phi)$ or $(\cos\phi, -\sin\phi)$. Plug p, q, r, t back in the third euqation in 1, we'll get

$$M = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix} \text{ or } \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

$\square$

(iii) False.

Solution 1: Assume $\exists C\,(A = CC^T)$, then for vector $\mathbf{x} = [1, 0, 0]$, we have

$$\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T CC^T\mathbf{x} \Rightarrow -8 = \langle C^T\mathbf{x}, C^T\mathbf{x}\rangle \Rightarrow -8 \geq 0.$$

The above inequality is invalid for all C, thus $\mathbf{x}$ is a counterexample to the statement.

Solution 2: Note that if a matrix can be written as $A = CC^T$, then $A^T = (CC^T)^T = CC^T = A$, *i.e.*, A is symmetric. However, the matrix $\begin{bmatrix} -8 & -1 & -6 \\ -3 & -5 & -7 \\ -4 & -9 & -2 \end{bmatrix}$ is not symmetric, thus the statement is false.

Solution 3: Assume C satisfies $A = CC^T$, $C \in \mathbb{R}^{3\times n}$. The entry of C at i^{th} row and j^{th} column is $C_{i.j}$. Express $A_{1,1}$ in terms of entries of C, we get $A_{1,1} = \sum_{j=1}^{n} C_{1,j}^2 \geq 0$. However, $A_{1,1} = -8$, which contradicts our assumption. Thus, A cannot be written as CC^T for any C.

2) Probability

(a)

(i)

Proof. According to Bayes' theorem,

$$\mathbb{E}[X] = \iint\limits_{X,Y} xp(x, y)\,dx\,dy = \int\limits_{Y} \left(\int\limits_{X} xp(x|y)\,dx\right) p(y)\,dy = \mathbb{E}_Y[\mathbb{E}_X[X|Y]].$$

□

(ii)

Proof.

$$\mathbb{E}[I|X \in \mathcal{C}] = \int\limits_{X} I[X \in \mathcal{C}]p(x)\,dx = \int\limits_{X\in\mathcal{C}} p(x)\,dx = P(X \in \mathcal{C})$$

□

(iii)

Proof.

$$\mathbb{E}_Y[\text{var}_X[X|Y]] = \mathbb{E}_Y[\mathbb{E}_X[X^2|Y]] - \mathbb{E}_Y[(\mathbb{E}_X[X|Y])^2]$$
$$\text{var}_Y[\mathbb{E}_X[X|Y]] = \mathbb{E}_Y[(\mathbb{E}_X[X|Y])^2] - \mathbb{E}_Y[(\mathbb{E}_X[X|Y])]^2$$

sum up the above two equations and use the results of (1), we get $\mathbb{E}_Y[\text{var}_X[X|Y]] + \text{var}_Y[\mathbb{E}_X[X|Y]] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \text{var}[X]$. □

(iv)

Proof. If X and Y are independent,

$$\mathbb{E}[XY] = \iint\limits_{X,Y} xyp(x,y)\,dx\,dy = \iint\limits_{X,Y} xyp(x)p(y)\,dx\,dy = \int\limits_X xp(x)\,dx \int\limits_Y yp(y)\,dy = \mathbb{E}[X]\mathbb{E}[Y].$$

□

(V)

Proof. Since X and Y can take values in {0,1} (note that, {0,1} is a set with two elements, not an interval),

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \Rightarrow P_{X,Y}(x=1, y=1) = P_X(x=1)P_Y(y=1).$$

Thus

$$\begin{aligned} P_{X,Y}(x=1, y=0) &= P_X(x=1) - P_{X,Y}(x=1, y=1) \\ &= P_X(x=1) - P_X(x=1)P_Y(y=1) \\ &= P_X(x=1)P_Y(y=0). \end{aligned}$$

Similarly, we can also get $P_{X,Y}(x=0, y=1) = P_X(x=0)P_Y(y=1)$ and $P_{X,Y}(x=0, y=0) = P_X(x=0)P_Y(y=0)$. So we can conclude that $P_{X,Y}(X,Y) = P_X(X)P_Y(Y), \forall X, Y \in \{0,1\} \Rightarrow X, Y$ are independent. □

(b)

(i) $\leq$. According to Bayes' theorem, we have $P(H=h, D=d) = P(D=d|H=h)P(H=h) \leq P(H=h)$, the inequality is due to the fact that $P(D=d|H=h) \leq 1$.

(ii) Depends. $P(H=h|D=d) = \frac{P(D=d|H=h)}{P(D=d)}P(H=h)$. If d and h have a large overlap, then $\frac{P(D=d|H=h)}{P(D=d)} > 1$, we have $P(H=h|D=d) > P(H=h)$; otherwise $P(H=h|D=d) \leq P(H=h)$.

(iii) $\geq$. $P(H=h|D=d) = \frac{P(D=d|H=h)P(H=h)}{P(D=d)} \geq P(D=d|H=h)P(H=h)$.

3) Positive (Semi-)Definite Matrices

(a)

Proof. $\Rightarrow$. If $\lambda_i \geq 0$ for each i, then

$$\mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{d} \lambda_i \mathbf{x}^T \mathbf{u_i}\mathbf{u_i}^T\mathbf{x} = \sum_{i=1}^{d} \lambda_i (\mathbf{x}^T\mathbf{u_i})^2 \geq 0,$$

thus A is PSD.
$\Leftarrow$. If A is PSD,

$$\lambda_i = \lambda_i \mathbf{u}_i^T\mathbf{u}_i = \mathbf{u}_i^T(A\mathbf{u}_i) \geq 0, \forall i \in \{1, 2, \ldots, d\}.$$

So the statement is true. $\square$

(b)

Proof. $\Rightarrow$. If $\lambda_i > 0$ for each i, then for $\mathbf{x} \neq \mathbf{0}$

$$\mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{d} \lambda_i \mathbf{x}^T \mathbf{u_i}\mathbf{u_i}^T\mathbf{x} = \sum_{i=1}^{d} \lambda_i (\mathbf{x}^T\mathbf{u_i})^2 > 0,$$

thus A is PD.
$\Leftarrow$. Because A is symmetric, due to the *spectral theorem*, $\mathbf{u}_i \neq \mathbf{0}$ is always true, then if A is PD, we have

$$\lambda_i = \lambda_i \mathbf{u}_i^T\mathbf{u}_i = \mathbf{u}_i^T(A\mathbf{u}_i) > 0, \forall i \in \{1, 2, \ldots, d\}.$$

So the statement is true. $\square$

4) Optimization

a)

Proof. For an affine function $f(t\mathbf{x} + (1-t)\mathbf{y}) = t\mathbf{a}^T\mathbf{x} + (1-t)\mathbf{a}^T\mathbf{y} + b = tf(\mathbf{x}) + (1-t)f(\mathbf{y})$. Thus, both $f(t\mathbf{x} + (1-t)\mathbf{y}) \geq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$ and $f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$ hold for an affine function, it's convex and concave. $f(\mathbf{x})$ is not strictly convex. $\square$

b)

Proof. Assume both $\mathbf{x}^*$ and $\mathbf{x}^{**}$ are global optimizers for f, and the optimal value is $\mathcal{O}(f)$. Then for $t \in [0,1]$, we have

$$f(t\mathbf{x}^* + (1-t)\mathbf{x}^{**}) < t\mathcal{O}(f) + (1-t)\mathcal{O}(f) = \mathcal{O}(f),$$

thus $\mathbf{x}^*$ and $\mathbf{x}^{**}$ are not the global optimizers, which contradicts our assumption. So a strict convex function has at most one global optimizer. □

c) With the first expansion, for any $\mathbf{y}$ we have

$$f(\mathbf{x}^* + t\mathbf{y}) = f(\mathbf{x}^*) + \langle \nabla f(\mathbf{x}^*), t\mathbf{y}\rangle + \frac{t^2}{2}\langle \mathbf{y}, \nabla^2 f(\mathbf{x}^*)\mathbf{y}\rangle + o(t^2\|\mathbf{y}\|^2).$$

Rearrange and note that $\nabla f(\mathbf{x}^*) = 0$, for sufficiently small t, we get

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

the inequality follows from the local optimality of x^*. Then take the limit on both sides, we get

$$\begin{aligned}\lim_{t\to 0}\frac{o(t^2\|\mathbf{y}\|^2)}{t^2\|\mathbf{y}\|^2} + \frac{1}{2\|\mathbf{y}\|^2}\langle \mathbf{y}, \nabla^2 f(\mathbf{x}^*)\mathbf{y}\rangle &\geq 0\\ \frac{1}{2\|\mathbf{y}\|^2}\langle \mathbf{y}, \nabla^2 f(\mathbf{x}^*)\mathbf{y}\rangle &\geq 0,\end{aligned}$$

thus, the Hessian is PSD.

d)

Proof. $\Rightarrow$. Assume the Hessian is PSD, then

$$\begin{aligned}f(\mathbf{x}) &= f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{1}{2}\langle \mathbf{x}-\mathbf{y}, \nabla^2 f(\mathbf{y}+t(\mathbf{x}-\mathbf{y}))(\mathbf{x}-\mathbf{y})\rangle\\ &\geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle.\end{aligned}$$

f is a convex function.
$\Leftarrow$. Assume f is convex, then for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

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

thus $\frac{1}{2}\langle \mathbf{y}, \nabla^2 f(\mathbf{x})\mathbf{y}\rangle + \frac{o(t^2\|\mathbf{y}\|^2)}{t^2} \geq 0 \Rightarrow \frac{1}{2}\langle \mathbf{y}, \nabla^2 f(\mathbf{x})\mathbf{y}\rangle \geq 0$, for sufficiently small t. Because $\mathbf{x}, \mathbf{y}$ are both arbitrary, the Hessian of f is PSD for all $X \in \mathbb{R}^d$. $\square$

e) We can express function f as

$$f(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} A_{i,j}x_i x_j + \sum_{i=1}^{d} b_i x_i + c.$$

Take the twice derivative of f, we get the $(i,j)^{th}$ entry of the Hessian matrix is

$$\nabla^2 f(\mathbf{x})_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j} = A_{i,j},$$

thus the Hessian of f is A. From the results of (d) we know, f is convex iff A is PSD. It's easy to show that f is strictly convex iff A is PD.
(**Important note**: you will encounter the quadratic form $\mathbf{x}^T A\mathbf{x}$ a lot in the future, so it's very helpful to memorize the results you derived: $\nabla^2 \mathbf{x}^T A\mathbf{x} = 2A$, and $\nabla^2 \mathbf{b}X = 0$. Actually these are basic results from *matrix derivatives*. If you are already familiar with matrix derivatives, you can directly use the results to solve this problem instead of writing the function f as summations.)
(**Note to graders**: directly using matrix derivatives is also a correct solution.)

5) Programming

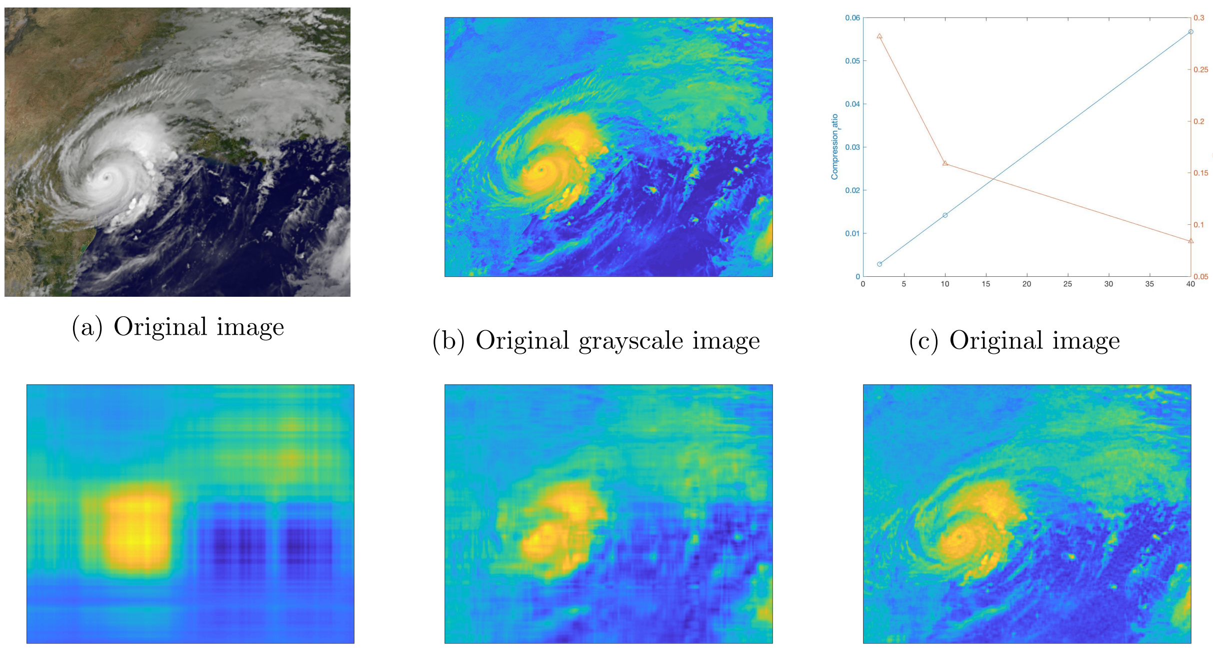
a) With MATLAB, the relative errors are {0.2815, 0.1587, 0.0837} for k = {2, 10, 40}. With Python, the relative errors are {0.2826, 0.1593, 0.0841}. The discrepancy is probably due to the fact that the SVD solver of Numpy package is less accurate than MATLAB's svd fuction.

The rank-k approximated images obtained from MATLAB are shown below. As k increases, the quality of the approximation improves.

b) The numbers required to describe the approximation are : (1) selected singular values + (2) numbers in those corresponding u_i + (3) numbers in those corresponding v_i. And the number of required numbers are {5690, 28450, 113800} for $k = \{2, 10, 40\}$.

Example code for problem 5

You can find the MATLAB and Python example code in Canvas.



(a) Original image

(b) Original grayscale image

(c) Original image

(d) $k = 2$

(e) $k = 10$

(f) $k = 40$

Figure 1: Compressed images