Dongyuan Wu.
08/04/2020.

Machine Learning Foundations : HW0

1. Probability and Statistics

(combinatorics)

(1). Let $C(N,K)=1$ for $K=0$ or $K=N$, and $C(N,K)=C(N-1,K)+C(N-1,K-1)$ for $N\geq 1$. Prove that $C(N,K)=\frac{N!}{K!(N-K)!}$ for $N\geq 1$ and $0\leq K\leq N$.

Proof: ① For $N=1$, K may only be 0 or 1. Then $C_N^K=\frac{1!}{0!1!}=1$ satisfies the statement.

② Let us assume the statement is true for $N=m$ $(m=2,3,\cdots)$, $0\leq K\leq N$.

Hence, $C_m^k=\frac{m!}{K!(m-K)!}$ is true. (It is an assumption.)

We have to prove that $C_{m+1}^K=\frac{(m+1)!}{K!(m+1-K)!}$ also holds.

$$C_{m+1}^K=C_m^K+C_m^{K-1}=\frac{m!}{K!(m-K)!}+\frac{m!}{(K-1)!(m-K+1)!}=\frac{m!(K-1)!(m-K+1)!+m!K!(m-K)!}{K!(m-K)!(K-1)!(m-K+1)!}$$

$$=\frac{m!(K-1)!(m-K)![K+(m-K+1)]}{K!(m-K)!(K-1)!(m-K+1)!}=\frac{m!(m+1)}{K!(m-K+1)!}=\frac{(m+1)!}{K!(m+1-K)!}$$

∴ The statement is also proved for $N=m+1$.

Hence, $C_N^K=\frac{N!}{K!(N-K)!}$ is proved for $N\geq 1$ and $0\leq K\leq N$.

(Counting)

(2) What is the probability of getting exactly 4 heads when flipping 10 fair coins? (a) What is the probability of getting a full house (XXXYY) when randomly draw 5 cards out of a deck of 52 cards? (b)

Solution: a). $P=\frac{C_{10}^4}{2^{10}}$ b). $P=\frac{C_{13}^1C_4^3C_{12}^1C_4^2}{C_{52}^5}$

(conditional probability)

(3) If your friend flipped a fair coin three times, and tell you that one of the tosses resulted in head, what is the probability that all three tosses resulted in heads?

Solution: Suppose event $A=\{$At least one of the tosses resulted in head$\}$,

event $B=\{$All the three tosses resulted in heads$\}$

Then $P(A)=1-\frac{1}{2^3}=\frac{7}{8}$, $P(B)=\frac{1}{2^3}=\frac{1}{8}$.

$\because P(A|B)=1$

$\therefore P(B|A)=\frac{P(A,B)}{P(A)}=\frac{P(A|B)P(B)}{P(A)}=\frac{1\cdot\frac{1}{8}}{\frac{7}{8}}=\frac{1}{7}$

(Bayes theorem)

(4) A program selects a random integer X like this. a random bit is first generated uniformly. If the bit is 0, X is drawn uniformly from $\{0,1,\cdots,7\}$; otherwise, X is drawn uniformly from $\{0,-1,-2,-3\}$. If we get an X from the program with $|X|=1$, what is the probability that X is negative?

Solution: Suppose event $A = \{$Get an X with $|X|=1\}$,

event $B = \{X$ is negative (i.e.: $-1)\}$.

Then $P(A) = \frac{1}{2}\times\frac{1}{8} + \frac{1}{2}\times\frac{1}{4} = \frac{3}{16}$, $P(B) = \frac{1}{2}\times\frac{1}{4} = \frac{1}{8}$

$\because P(A|B) = 1$

$\therefore P(B|A) = \frac{P(A,B)}{P(A)} = \frac{P(A|B)\cdot P(B)}{P(A)} = \frac{1\cdot\frac{1}{8}}{\frac{3}{16}} = \frac{2}{3}$

(union/intersection)

(5). If $P(A)=0.3$ and $P(B)=0.4$, a). what is the maximum possible value of $P(A\cap B)$? b). What is the minimum possible value of $P(A\cap B)$? c). What is the maximum possible value of $P(A\cup B)$? d). What is the minimum possible value of $P(A\cup B)$?

Solution: a). 0.3 b). 0 c). 0.7 d). 0.4

2. Linear Algebra

(rank)

(4). What is the rank of $\begin{pmatrix}1&2&1\\1&0&3\\1&1&2\end{pmatrix}$?

Solution: $\because A = \begin{pmatrix}1&2&1\\1&0&3\\1&1&2\end{pmatrix} \to \begin{pmatrix}1&2&1\\0&-2&2\\0&-1&1\end{pmatrix} \to \begin{pmatrix}1&2&1\\0&-2&2\\0&0&0\end{pmatrix}$

$\therefore \text{rank}(A) = 2$

(inverse)

(2). What is the inverse of $\begin{pmatrix}0&2&4\\2&4&2\\3&3&1\end{pmatrix}$?

Solution: $\because [A\ I] = \begin{pmatrix}0&2&4&1&0&0\\2&4&2&0&1&0\\3&3&1&0&0&1\end{pmatrix} \to \begin{pmatrix}0&1&2&\frac{1}{2}&0&0\\1&2&1&0&\frac{1}{2}&0\\3&3&1&0&0&1\end{pmatrix} \to \begin{pmatrix}1&2&1&0&\frac{1}{2}&0\\0&1&2&\frac{1}{2}&0&0\\0&-3&-2&0&-\frac{3}{2}&1\end{pmatrix}$

$\to \begin{pmatrix}1&0&-3&-1&\frac{1}{2}&0\\0&1&2&\frac{1}{2}&0&0\\0&0&4&\frac{3}{2}&-\frac{3}{2}&1\end{pmatrix} \to \begin{pmatrix}1&0&-3&-1&\frac{1}{2}&0\\0&1&2&\frac{1}{2}&0&0\\0&0&1&\frac{3}{8}&-\frac{3}{8}&\frac{1}{4}\end{pmatrix} \to \begin{pmatrix}1&0&0&\frac{1}{8}&-\frac{5}{8}&\frac{3}{4}\\0&1&0&-\frac{1}{4}&\frac{3}{4}&-\frac{1}{2}\\0&0&1&\frac{3}{8}&-\frac{3}{8}&\frac{1}{4}\end{pmatrix}$

$\therefore A^{-1} = \begin{pmatrix}\frac{1}{8}&-\frac{5}{8}&\frac{3}{4}\\-\frac{1}{4}&\frac{3}{4}&-\frac{1}{2}\\\frac{3}{8}&-\frac{3}{8}&\frac{1}{4}\end{pmatrix}$

(eigenvalues/eigenvectors)

(3). What are the eigenvalues and eigenvectors of $\begin{pmatrix}3&1&1\\2&4&2\\-1&-1&1\end{pmatrix}$?

Solution: $|A-\lambda I| = \begin{vmatrix} 3-\lambda & 1 & 1 \\ 2 & 4-\lambda & 2 \\ -1 & -1 & 1-\lambda \end{vmatrix} = \begin{vmatrix} 3-\lambda & 1 & 1 \\ 0 & 2-\lambda & 4-2\lambda \\ -1 & -1 & 1-\lambda \end{vmatrix} = \begin{vmatrix} 0 & \lambda-2 & (1-\lambda)(3-\lambda)+1 \\ 0 & 2-\lambda & 4-2\lambda \\ -1 & -1 & 1-\lambda \end{vmatrix}$

$= \begin{vmatrix} 1 & 1 & \lambda-1 \\ 0 & 2-\lambda & 4-2\lambda \\ 0 & \lambda-2 & (1-\lambda)(3-\lambda)+1 \end{vmatrix} = \begin{vmatrix} 1 & 1 & \lambda-1 \\ 0 & 2-\lambda & 4-2\lambda \\ 0 & 0 & (1-\lambda)(3-\lambda)+1+4-2\lambda \end{vmatrix}$

$= \begin{vmatrix} 1 & 1 & \lambda-1 \\ 0 & 2-\lambda & 4-2\lambda \\ 0 & 0 & (\lambda-2)(\lambda-4) \end{vmatrix} = -(\lambda-2)^2(\lambda-4)$

Suppose $|A-\lambda I|=0$, then the eigenvalues of A are $\lambda_1=\lambda_2=2$ and $\lambda_3=4$.

For $\lambda_1=\lambda_2=2$, $\begin{pmatrix} 3 & 1 & 1 \\ 2 & 4 & 2 \\ -1 & -1 & 1 \end{pmatrix}\begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \end{pmatrix} = 2\begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \end{pmatrix} \Rightarrow x_{11}+x_{21}+x_{31}=0$

$\therefore$ The corresponding eigenvectors ~~are~~ could be $x_1 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$ and $x_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$.

For $\lambda_3=4$, $\begin{pmatrix} 3 & 1 & 1 \\ 2 & 4 & 2 \\ -1 & -1 & 1 \end{pmatrix}\begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \end{pmatrix} = 4\begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \end{pmatrix} \Rightarrow \begin{cases} x_{11} = -x_{31} \\ x_{21} = -2x_{31} \end{cases}$

$\therefore$ The corresponding ~~eigenvalue~~ eigenvector ~~is~~ could be $x_3 = \begin{pmatrix} 1 \\ 2 \\ -1 \end{pmatrix}$.

(singular value decomposition)

(4). a). For a real matrix M, let $M=U\Sigma V^T$ be its singular value decomposition. Define $M^\dagger = V\Sigma^\dagger U^T$, where $\Sigma^\dagger[i][j] = \frac{1}{\Sigma[i][j]}$ when $\Sigma[i][j]$ is nonzero, and 0 otherwise. Prove that $MM^\dagger M = M$.

b). If M is invertible, prove that $M^\dagger = M^{-1}$.

Proof: a). $\because M=U\Sigma V^T$ is the singular value decomposition

$\therefore U$ and V are both the orthonormal matrix, i.e. $UU^T=VV^T=I=U^TU=V^TV$

$\therefore MM^\dagger M = U\Sigma V^T V\Sigma^\dagger U^T U\Sigma V^T = U\Sigma\Sigma^\dagger\Sigma V^T$

$\because \Sigma^\dagger[i][j] = \frac{1}{\Sigma[i][j]}$ when $\Sigma[i][j]$ is nonzero, and 0 otherwise.

$\therefore \Sigma^\dagger\Sigma = I$

$\therefore MM^\dagger M = U\Sigma\Sigma^\dagger\Sigma V^T = U\Sigma V^T = M$

b). $\because M$ is invertible

$\therefore MM^{-1} = M^{-1}M = I$

$\because MM^\dagger M = M$

$\therefore M^{-1}MM^\dagger MM^{-1} = M^{-1}MM^{-1} \Rightarrow M^\dagger = M^{-1}$

(PD/PSD)

(5). A symmetric real matrix A is positive definite (PD) iff. $x^TAx>0$ for all $x\neq 0$, and positive semi-definite (PSD) if ">" is changed to "$\geq$". Prove:

a). For any real matrix Z, ZZ^T is PSD.

b). A symmetric A is PD iff. all eigenvalues of A are strictly positive.

Proof: a). For any x, $x^TZZ^Tx=(Z^Tx)^T(Z^Tx)=\|Z^Tx\|^2\geq 0$

$\therefore ZZ^T\geq 0$ i.e. ZZ^T is PSD. $((ZZ^T)^T=ZZ^T \Rightarrow ZZ^T$ is symmetric).

b). ($\Rightarrow$): ~~For $\forall x\neq 0$, let $y=T^T$~~

$\because A$ is a symmetric matrix

$\therefore \exists$ an orthonormal matrix T s.t. $T^TAT=\Lambda=diag(\lambda_1,\lambda_2,\cdots\lambda_p)$.

For $\forall x\neq 0$, let $y=T^Tx$, then $x^TAx=x^TT\Lambda T^Tx=y^T\Lambda y=\sum_{i=1}^p\lambda_iy_i^2$

$\because y\neq 0$ i.e. at least one of $y_1,y_2,\cdots,y_p$ are nonzero

$\therefore$ If $\lambda_i>0$ $(i=1,2,\cdots,p)$, then $x^TAx=\sum_{i=1}^p\lambda_iy_i^2>0$

$\therefore A$ is PD.

($\Leftarrow$): Assume $e_i=\begin{pmatrix}0\\\vdots\\0\\1\\0\\\vdots\\0\end{pmatrix}$, "1" is at the i-th position. Then we have

$\lambda_i=e_i^T\Lambda e_i=e_i^TT^TATe_i=(Te_i)^TA(Te_i)$

$\because Te_i\neq 0$

$\therefore$ If A is PD, i.e. $A>0$, then $\lambda_i>0$ $(i=1,2,\cdots,p)$.

(inner product)

b) Consider $x\in R^d$ and some $u\in R^d$ with $\|u\|=1$. a). What is the maximum value of u^Tx? What u results in the maximum value? b). What is the minimum value of u^Tx? What u results in the minimum value? c). What is the minimum value of $|u^Tx|$? What u results in the minimum value?

Solution: a). ~~$\max(u^Tx)=\|x\|$ $u=$~~ $u^Tx=\|u\|\cdot\|x\|\cdot\cos\theta=\|x\|\cdot\cos\theta$

$\because \cos\theta\in[-1,1]$ $\therefore \max(u^Tx)=\|x\| \Rightarrow u=\frac{x}{\|x\|}$

b). $\min(u^Tx)=-\|x\| \Rightarrow u=-\frac{x}{\|x\|}$

c). $\min(|u^Tx|)=0$ (when $\theta=90^\circ$, $\cos\theta=0$). $\Rightarrow u\perp x$.

3. Calculus

(differential and partial differential)

(1) Let $f(x)=\ln(1+e^{-2x})$. What is $\frac{df(x)}{dx}$? Let $g(x,y)=e^x+e^{2y}+e^{3xy^2}$. What is $\frac{\partial g(x,y)}{\partial y}$?

Solution: $\frac{df(x)}{dx}=\frac{1}{1+e^{-2x}}\cdot e^{-2x}\cdot(-2)=-\frac{2e^{-2x}}{1+e^{-2x}}$

$\frac{\partial g(x,y)}{\partial y}=0+e^{2y}\cdot 2+e^{3xy^2}\cdot 6xy=2e^{2y}+6xye^{3xy^2}$.

(chain rule)

(2) Let $f(x,y)=xy$, $x(u,v)=\cos(u+v)$, $y(u,v)=\sin(u-v)$. What is $\frac{\partial f}{\partial v}$?

Solution: $\frac{\partial f}{\partial v}=y\cdot\frac{\partial x}{\partial v}+x\cdot\frac{\partial y}{\partial v}=\sin(u-v)[-\sin(u+v)]+\cos(u+v)[-\cos(u-v)]$

$=-[\sin(u-v)\sin(u+v)+\cos(u-v)\cos(u+v)]=-\cos[(u+v)-(u-v)]$

$=-\cos(2v)$

(gradient and Hessian)

(3) Let $E(u,v)=(ue^v-2ve^{-u})^2$. Calculate the gradient ∇E and the Hessian $\nabla^2 E$ at $u=1$ and $v=1$.

Solution: $\frac{\partial E}{\partial u}=2(ue^v-2ve^{-u})(e^v+2ve^{-u})$; $\frac{\partial E}{\partial v}=2(ue^v-2ve^{-u})(ue^v-2e^{-u})$

$\frac{\partial E}{\partial u}=2ue^{2v}+4uve^{v-u}-4ve^{v-u}-8ve^{-2u}$; $=2u^2e^{2v}-4ue^{v-u}-4uve^{v-u}+8ve^{-2u}$

$\frac{\partial^2 E}{\partial u^2}=2e^{2v}+4ve^{v-u}-4uve^{v-u}+4ve^{v-u}+16v^2e^{-2u}$

$\frac{\partial^2 E}{\partial u\partial v}=4ue^{2v}+4ue^{v-u}+4uve^{v-u}-4e^{v-u}-4ve^{v-u}-16ve^{-2u}$

$\frac{\partial^2 E}{\partial v\partial u}=4ue^{2v}-4e^{v-u}+4ue^{v-u}-4ve^{v-u}+4uve^{v-u}-16ve^{-2u}$

$\frac{\partial^2 E}{\partial v^2}=4u^2e^{2v}-4ue^{v-u}-4ue^{v-u}-4uve^{v-u}+8e^{-2u}$

$\therefore \nabla E|_{u=1,v=1}=\left[\frac{\partial E}{\partial u},\frac{\partial E}{\partial v}\right]\Big|_{u=1,v=1}=(2e^2+4-4-8e^{-2},\ 2e^2-4-4+8e^{-2})=(2e^2-8e^{-2},\ 2e^2+8e^{-2})$

$\nabla^2 E|_{u=1,v=1}=\begin{bmatrix}\frac{\partial^2 E}{\partial u^2} & \frac{\partial^2 E}{\partial u\partial v}\\ \frac{\partial^2 E}{\partial v\partial u} & \frac{\partial^2 E}{\partial v^2}\end{bmatrix}\Big|_{u=1,v=1}=\begin{bmatrix}2e^2+4-4+4+16e^{-2} & 4e^2+4+4-4-4-16e^{-2}\\ 4e^2-4+4-4+4-16e^{-2} & 4e^2-4-4-4+8e^{-2}\end{bmatrix}$

$=\begin{bmatrix}2e^2+16e^{-2}+4 & 4e^2-16e^{-2}\\ 4e^2-16e^{-2} & 4e^2+8e^{-2}-12\end{bmatrix}$

(Taylor's expansion)

(4) Let $E(u,v)=(ue^v-2ve^{-u})^2$. Write down the second-order Taylor's expansion of E around $u=1$ and $v=1$

Solution: $E(1,1)=(e-2e^{-1})^2=e^2+4e^{-2}-4$, $E_u(1,1)=2e^2-8e^{-2}$, $E_v(1,1)=2e^2+8e^{-2}-8$

$E_{uu}(1,1)=2e^2+16e^{-2}+4$, $E_{vv}(1,1)=4e^2+8e^{-2}-12$, $E_{uv}(1,1)=4e^2-16e^{-2}$

$\therefore E(u,v)\approx E(1,1)+E_u(1,1)(u-1)+E_v(1,1)(v-1)+\frac{1}{2}E_{uu}(1,1)(u-1)^2+E_{uv}(1,1)(u-1)(v-1)+\frac{1}{2}E_{vv}(1,1)(v-1)^2$

$=e^2+4e^{-2}-4+(2e^2-8e^{-2})(u-1)+(2e^2+8e^{-2}-8)(v-1)+(e^2+8e^{-2}+2)(u-1)^2+(2e^2+4e^{-2}-6)(v-1)^2+(4e^2-16e^{-2})$
$(u-1)(v-1)$

(optimization) (5). For some given $A>0, B>0$, solve $\min_{\alpha} Ae^{\alpha}+Be^{-2\alpha}$.

Solution: Let $f(\alpha)=Ae^{\alpha}+Be^{-2\alpha}$, then $f'(\alpha)=Ae^{\alpha}-2Be^{-2\alpha}$, $f''(\alpha)=Ae^{\alpha}+4Be^{-2\alpha}$

Assume $f'(\alpha)=0 \Rightarrow Ae^{\alpha}-2Be^{-2\alpha}=0 \Rightarrow e^{3\alpha}=\frac{2B}{A} \Rightarrow \alpha=\frac{1}{3}\ln\frac{2B}{A}$.

$\because f''(\frac{1}{3}\ln\frac{2B}{A})=A(\frac{2B}{A})^{\frac{1}{3}}+4B\cdot(\frac{2B}{A})^{-\frac{2}{3}}=2^{\frac{1}{3}}B^{\frac{1}{3}}A^{\frac{2}{3}}+2^{\frac{2}{3}}B^{\frac{1}{3}}A^{\frac{2}{3}}=3B^{\frac{1}{3}}A^{\frac{2}{3}}\cdot(2^{\frac{1}{3}}+2^{\frac{2}{3}})=3\cdot 2^{\frac{1}{3}}A^{\frac{2}{3}}B^{\frac{1}{3}}$

$\because A>0$ and $B>0$

$\therefore f''(\frac{1}{3}\ln\frac{2B}{A})=3\cdot 2^{\frac{1}{3}}\cdot A^{\frac{2}{3}}\cdot B^{\frac{1}{3}}>0$

$\therefore \alpha=\frac{1}{3}\ln\frac{2B}{A}$ is the point can get the minimum value.

$\therefore \min_{\alpha} Ae^{\alpha}+Be^{-2\alpha}=\frac{1}{3}\ln\frac{2B}{A}$.

(vector calculus) (6). Let w be a vector in R^d and $E(w)=\frac{1}{2}w^TAw+b^Tw$ for some symmetric matrix A and vector b.

Prove that the gradient $\nabla E(w)=Aw+b$ and the Hessian $\nabla^2 E(w)=A$.

Proof

Solution: $\because A$ is symmetric

$\therefore A$ is a square matrix

Assume $w=\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$, $A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1d} \\ a_{21} & a_{22} & \cdots & a_{2d} \\ \vdots & \vdots & & \vdots \\ a_{d1} & a_{d2} & \cdots & a_{dd} \end{pmatrix}$, $b=\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{pmatrix}$. Then we have:

$E(w)=\frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}a_{ij}w_iw_j+\sum_{j=1}^{d}b_jw_j=\frac{1}{2}\sum_{j=1}^{d}\sum_{i=1}^{d}a_{ij}w_j^2+\sum_{j=1}^{d}b_jw_j$

$\because \frac{\partial}{\partial w_k}[E(w)]=w_k\cdot a_{ij}+b_k$, $\frac{\partial^2}{\partial w_k^2}[E(w)]=a_{ij}$

$\therefore \nabla E(w)=Aw+b$, $\nabla^2 E(w)=A$.