大家好，这篇是有关台大机器学习课程作业零的详解。

我的github地址：
https://github.com/Doraemonzzz

个人主页：
http://doraemonzzz.com/

作业地址:
https://www.csie.ntu.edu.tw/~htlin/course/ml15fall/

参考资料:
https://blog.csdn.net/a1015553840/article/details/51085129
http://www.vynguyen.net/category/study/machine-learning/page/6/
http://book.caltech.edu/bookforum/index.php
http://beader.me/mlnotebook/ https://acecoooool.github.io/blog/

1 Probability and Statistics

(1) (combinatorics)

构造如下命题：

$$P(N)=\{C(N,K)=\frac{N!}{K!(N-K)!},0\le K\le N\}$$

对上述命题关于N做数学归纳法。

当$N=0$时，

$$C(N,K)=C(0,0)=\frac{0!}{0!0!}=1$$

所以$N=0$时结论成立。

假设$N=n$时结论成立，现在将推出$N=n+1$时结论也成立。事实上，对$K=n+1$，

$$C(n+1,K)=C(n+1,n+1)=1$$

对$K=0$，

$$C(n+1,K)=C(n+1,0)=1$$

对$1\le K\le n$，我们有

$$
\begin{aligned}
C(n+1,K) &= C(n,K)+C(n,K-1) \\
&= \frac{n!}{K!(n-K)!}+\frac{n!}{(K-1)!(n-K+1)!} \\
&= \frac{n!}{K!(n-K+1)!}(n-K+1+K) \\
&= \frac{n!}{K!(n-K+1)!}(n+1) \\
&= \frac{(n+1)!}{K!(n-K+1)!}
\end{aligned}
$$

所以$N=n+1$时结论也成立，原结论得证。

(2) (counting)

概率1：

$$
p_1=\frac{C_{10}^4}{2^{10}}=\frac{105}{512}
$$

概率2：

$$
p_2=\frac{13\times 12\times C_4^3\times C_4^2}{C_{52}^5}=\frac{6}{4165}
$$

(3) (conditional probability)

记

$$
\begin{aligned}
A &= \{\text{三次抛硬币的结果中有一次正面朝上}\} \\
B &= \{\text{三次抛硬币的结果中有三次正面朝上}\}
\end{aligned}
$$

那么

$$
\begin{aligned}
p &= \mathbb{P}(B|A) \\
&= \frac{\mathbb{P}(AB)}{\mathbb{P}(A)} \\
&= \frac{1/8}{1-1/8} \\
&= \frac{1}{7}
\end{aligned}
$$

(4) (Bayes theorem)

记

$$
\begin{aligned}
A &= \{|X| = 1\} \\
B &= \{X < 0\}
\end{aligned}
$$

那么

$$
\begin{aligned}
p &= \mathbb{P}(B|A) \\
&= \frac{\mathbb{P}(AB)}{\mathbb{P}(A)} \\
&= \frac{\mathbb{P}(X = -1)}{\mathbb{P}(|X| = 1)} \\
&= \frac{\frac{1}{2} \times \frac{1}{4}}{\frac{1}{2} \times \frac{1}{8} + \frac{1}{2} \times \frac{1}{4}} \\
&= \frac{2}{3}
\end{aligned}
$$

(5) (union/intersection)

首先显然有

$$
\begin{aligned}
\min \mathbb{P}(A \cap B) &= 0 \\
\max \mathbb{P}(A \cap B) &= \min\{\mathbb{P}(A), \mathbb{P}(B)\} = 0.3
\end{aligned}
$$

接着由容斥原理，我们有

$$
\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = 0.7 - \mathbb{P}(A \cap B)
$$

所以

$$
\begin{aligned}
\max \mathbb{P}(A \cup B) &= 0.7 \\
\min \mathbb{P}(A \cup B) &= 0.4
\end{aligned}
$$

(6) (mean/variance)

展开即可

$$
\begin{aligned}
\sigma_X^2 &= \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \overline{X})^2 \\
&= \frac{1}{N-1}\left(\sum_{n=1}^{N} X_n^2 - 2\sum_{n=1}^{N} X_n \overline{X} + \sum_{n=1}^{N} \overline{X}^2\right) \\
&= \frac{1}{N-1}\left(\sum_{n=1}^{N} X_n^2 - 2N\overline{X}^2 + N\overline{X}^2\right) \\
&= \frac{1}{N-1}\left(\sum_{n=1}^{N} X_n^2 - N\overline{X}^2\right) \\
&= \frac{N}{N-1}\left(\frac{1}{N}\sum_{n=1}^{N} X_n^2 - \overline{X}^2\right)
\end{aligned}
$$

(7) (Gaussian distribution)

由高斯分布的性质可得

$$
Z = X_1 + X_2
$$

也为高斯分布。（证明方法可以使用特征函数，这里从略）

接着计算期望和方差（利用独立性）：

$$
\begin{aligned}
\mathbb{E}[Z] &= \mathbb{E}[X_1 + X_2] \\
&= \mathbb{E}[X_1] + \mathbb{E}[X_2] \\
&= 2 - 3 \\
&= -1 \\
\mathrm{Var}(Z) &= \mathrm{Var}(X_1 + X_2) \\
&= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) \\
&= 1 + 4 \\
&= 5
\end{aligned}
$$

2 Linear Algebra

(1) (rank)

做初等行变化：

$$
\begin{pmatrix} 1 & 2 & 1 \\ 1 & 0 & 3 \\ 1 & 1 & 2 \end{pmatrix} \xrightarrow{(2)-(1)} \begin{pmatrix} 1 & 2 & 1 \\ 0 & -2 & 2 \\ 1 & 1 & 2 \end{pmatrix} \xrightarrow{(3)-(1)} \\
\begin{pmatrix} 1 & 2 & 1 \\ 0 & -2 & 2 \\ 0 & -1 & 1 \end{pmatrix} \xrightarrow{(2)/2} \begin{pmatrix} 1 & 2 & 1 \\ 0 & -1 & 1 \\ 0 & -1 & 1 \end{pmatrix} \xrightarrow{(3)+(2)} \\
\begin{pmatrix} 1 & 2 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{pmatrix}
$$

所以rank为2。

(2) (inverse)

记

$$
A = \begin{pmatrix} 0 & 2 & 4 \\ 2 & 4 & 2 \\ 3 & 3 & 1 \end{pmatrix}
$$

那么

$$
|A| = -16
$$

伴随矩阵为

$$
A^* = \begin{pmatrix} -2 & 10 & -12 \\ 4 & -12 & 8 \\ -6 & 6 & -4 \end{pmatrix}
$$

$$
A^{-1} = -\frac{1}{16} \begin{pmatrix} -2 & 10 & -12 \\ 4 & -12 & 8 \\ -6 & 6 & -4 \end{pmatrix}
$$

(3) (eigenvalues/eigenvectors)

记

$$
A = \begin{pmatrix} 3 & 1 & 1 \\ 2 & 4 & 2 \\ -1 & -1 & 1 \end{pmatrix}
$$

那么

$$(A-\lambda I)=\begin{pmatrix}3-\lambda & 1 & 1\\ 2 & 4-\lambda & 2\\ -1 & -1 & 1-\lambda\end{pmatrix}\xrightarrow{(1)+(2)+(3)}\begin{pmatrix}4-\lambda & 4-\lambda & 4-\lambda\\ 2 & 4-\lambda & 2\\ -1 & -1 & 1-\lambda\end{pmatrix}\xrightarrow{(2)-\frac{2}{4-\lambda}\times(1)}$$
$$\begin{pmatrix}4-\lambda & 4-\lambda & 4-\lambda\\ 0 & 2-\lambda & 0\\ -1 & -1 & 1-\lambda\end{pmatrix}\xrightarrow{(3)+\frac{1}{4-\lambda}(1)}\begin{pmatrix}4-\lambda & 4-\lambda & 4-\lambda\\ 0 & 2-\lambda & 0\\ 0 & 0 & 2-\lambda\end{pmatrix}$$

所以特征多项式为

$$|A-\lambda I|=(4-\lambda)(2-\lambda)^2$$

特征值为

$$\lambda_1=4,\lambda_2=\lambda_3=2$$

接着求特征向量，当$\lambda=4$时，

$$A-4I=\begin{pmatrix}-1 & 1 & 1\\ 2 & 0 & 2\\ -1 & -1 & -3\end{pmatrix}$$

求解

$$(A-4I)\vec{x}=0$$

可得特征向量为

$$\vec{x}_1=\begin{pmatrix}1\\ 2\\ -1\end{pmatrix}$$

当$\lambda=2$时，

$$A-2I=\begin{pmatrix}1 & 1 & 1\\ 2 & 2 & 2\\ -1 & -1 & -1\end{pmatrix}$$

求解

$$(A-2I)\vec{x}=0$$

可得特征向量为

$$\vec{x}_2=\begin{pmatrix}1\\ -1\\ 0\end{pmatrix},\vec{x}_3=\begin{pmatrix}1\\ 0\\ -1\end{pmatrix}$$

(4) (singular value decomposition)

(a)奇异值分解的形式如下：

$$
\begin{gathered}
M=U\Sigma V^T \\
\text{其中} M\in\mathbb{R}^{m\times n}, U\in\mathbb{R}^{m\times m}, \Sigma\in\mathbb{R}^{m\times n}, V\in\mathbb{R}^{n\times n} \\
UU^T=U^TU=I_m, VV^T=V^TV=I_n
\end{gathered}
$$

根据定义，这里应该有

$$
\Sigma^\dagger\in\mathbb{R}^{n\times m}
$$

所以

$$
\begin{aligned}
MM^\dagger M&=U\Sigma V^TV\Sigma^\dagger U^TU\Sigma V^T\\
&=U(\Sigma\Sigma^\dagger)\Sigma V^T\\
&=UI_m\Sigma V^T\\
&=U\Sigma V^T
\end{aligned}
$$

(b)如果M可逆，那么

$$
m=n
$$

所以

$$
\Sigma^\dagger=\Sigma^{-1}
$$

因此

$$
\begin{aligned}
M^\dagger&=V\Sigma^\dagger U^T\\
&=V\Sigma^{-1}U^T\\
&=(U\Sigma V^T)^{-1}\\
&=M^{-1}
\end{aligned}
$$

(5) (PD/PSD)

(a)$\forall x$，我们有

$$
x^TZZ^Tx=(Z^Tx)^T(Z^Tx)=\|Z^Tx\|_2^2\geq 0
$$

(b)因为对称矩阵正交相似于对角阵，所以存在正交矩阵Q，和对角阵Λ，使得

$$
Q^TAQ=\Lambda
$$

其中

$$
\Lambda=\operatorname{diag}\{\lambda_1,\ldots,\lambda_n\}
$$

$\Rightarrow$

取$x=Qe_i\neq 0, i=1,\ldots,n$，其中

$$
e_i\in\mathbb{R}^n, (e_i)_j=1\{i=j\}
$$

那么

$$
\begin{aligned}
x^T\Lambda x &= e_i^TQ^TAQe_i\\
&= e_i^T\Lambda e_i\\
&= \lambda_i\\
&> 0
\end{aligned}
$$

结论得证。

$\Leftarrow$

$\forall x$，令

$$
y = Q^Tx
$$

那么我们有

$$
\begin{aligned}
x^TAx &= x^TQ\Lambda Q^Tx\\
&= y^T\Lambda y\\
&= \sum_{i=1}^n \lambda_i y_i^2\\
&\geq 0
\end{aligned}
$$

因为$\lambda_i>0$，所以上式为0当且仅当

$$
y_i=0, i=1,\ldots,n
$$

即

$$
y = Q^Tx = 0
$$

左乘Q得到

$$
x = 0
$$

结论得证。

(6) (inner product)

(a)(b)

利用柯西不等式可得

$$
|u^Tx| \leq |u^T|.|x| = |x|
$$

所以

$$
-|x| \leq u^Tx \leq |x|
$$

即

$$
\begin{aligned}
\max u^T x &= |x| \qquad u,x\text{同 向 时 取 等 号} \\
\min u^T x &= -|x| \qquad u,x\text{反 向 时 取 等 号}
\end{aligned}
$$

(c)显然有

$$
\min |u^T x| = 0 \qquad u,x\text{正 交 时 取 等 号}
$$

(7) (distance)

$\forall x_1 \in H_1, x_2 \in H_2$，根据投影的定义，距离为

$$
\begin{aligned}
d &= \frac{|w^T(x_1 - x_2)|}{\|w\|} \\
&= |3+2| \\
&= 5
\end{aligned}
$$

3 Calculus

(1) (differential and partial differential)

$$
\begin{aligned}
\frac{df(x)}{dx} &= \frac{-2e^{-2x}}{1+e^{-2x}} \\
&= -\frac{2}{1+e^{2x}} \\
\frac{\partial g(x,y)}{\partial y} &= 2e^{2y} + e^{3xy^2} \times 6xy \\
&= 6xye^{3xy^2} + 2e^{2y}
\end{aligned}
$$

(2) (chain rule)

$$
\begin{aligned}
\frac{\partial f}{\partial v} &= \frac{\partial f}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial v} \\
&= y(-\sin(u+v)) + x(-\cos(u-v)) \\
&= -y\sin(u+v) - x\cos(u-v)
\end{aligned}
$$

(3) (integral)

$$
\begin{aligned}
\int_5^{10} \frac{2}{x-3} dx &= 2\ln(x-3)\Big|_5^{10} \\
&= 2\ln\frac{7}{2}
\end{aligned}
$$

(4) (gradient and Hessian)

首先求一阶偏导数:

$$
\begin{aligned}
\frac{\partial E(u,v)}{\partial u} &= 2(ue^{v} - 2ve^{-u})(e^{v} + 2ve^{-u}) \\
&= 2(ue^{2v} - 2ve^{v-u} + 2uve^{v-u} - 4v^2 e^{-2u}) \\
\frac{\partial E(u,v)}{\partial v} &= 2(ue^{v} - 2ve^{-u})(ue^{v} - 2e^{-u}) \\
&= 2(u^2 e^{2v} - 2uve^{v-u} - 2ue^{v-u} + 4ve^{-2u})
\end{aligned}
$$

接着求二阶偏导数:

$$
\begin{aligned}
\frac{\partial^2 E(u,v)}{\partial u^2} &= 2\frac{\partial}{\partial u}(ue^{2v} - 2ve^{v-u} + 2uve^{v-u} - 4v^2 e^{-2u}) \\
&= 2(e^{2v} + 2ve^{v-u} + 2ve^{v-u} - 2uve^{v-u} + 8v^2 e^{-2u}) \\
&= 2(e^{2v} + 4ve^{v-u} - 2uve^{v-u} + 8v^2 e^{-2u}) \\
\frac{\partial^2 E(u,v)}{\partial v^2} &= 2\frac{\partial}{\partial v}(u^2 e^{2v} - 2uve^{v-u} - 2ue^{v-u} + 4ve^{-2u}) \\
&= 2(2u^2 e^{2v} - 2ue^{v-u} - 2uve^{v-u} - 2ue^{v-u} + 4e^{-2u}) \\
&= 2(2u^2 e^{2v} - 4ue^{v-u} - 2uve^{v-u} + 4e^{-2u}) \\
\frac{\partial^2 E(u,v)}{\partial u \partial v} &= 2\frac{\partial}{\partial v}(ue^{2v} - 2ve^{v-u} + 2uve^{v-u} - 4v^2 e^{-2u}) \\
&= 2(2ue^{2v} - 2e^{v-u} - 2ve^{v-u} + 2ue^{v-u} + 2uve^{v-u} - 8ve^{-2u})
\end{aligned}
$$

将$u=v=1$，带入可得

$$
\begin{aligned}
\frac{\partial E(u,v)}{\partial u}\Big|_{u=1,v=1} &= 2(e^2 - 4e^{-2}) \\
\frac{\partial E(u,v)}{\partial v}\Big|_{u=1,v=1} &= 2(e^2 + 4e^{-2} - 4) \\
\frac{\partial^2 E(u,v)}{\partial u^2}\Big|_{u=1,v=1} &= 2(e^2 + 8e^{-2} + 2) \\
\frac{\partial^2 E(u,v)}{\partial v^2}\Big|_{u=1,v=1} &= 2(2e^2 + 4e^{-2} - 6) \\
\frac{\partial^2 E(u,v)}{\partial u \partial v}\Big|_{u=1,v=1} &= 2(2e^2 - 8e^{-2})
\end{aligned}
$$

因此

$$
\nabla E = \begin{pmatrix} 2(e^2 - 4e^{-2}) \\ 2(e^2 + 4e^{-2} - 4) \end{pmatrix}, \nabla^2 E = \begin{pmatrix} 2(e^2 + 8e^{-2} + 2) & 2(2e^2 - 8e^{-2}) \\ 2(2e^2 - 8e^{-2}) & 2(2e^2 + 4e^{-2} - 6) \end{pmatrix}
$$

(5) (Taylor's expansion)

泰勒展开可得

$$
\begin{aligned}
E(u,v) &\approx E(1,1)+\nabla E^T\begin{pmatrix}u-1\\ v-1\end{pmatrix}+\frac{1}{2}\begin{pmatrix}u-1 & v-1\end{pmatrix}\nabla^2 E\begin{pmatrix}u-1\\ v-1\end{pmatrix}\\
&=(e-2e^{-1})^2+2(e^2-4e^{-2})(u-1)+2(e^2+4e^{-2}-4)(v-1)\\
&+(e^2+8e^{-2}+2)(u-1)^2+(2e^2+4e^{-2}-6)(v-1)^2+2(2e^2-8e^{-2})(u-1)(v-1)
\end{aligned}
$$

(6) (optimization)

记

$$
f(\alpha)=Ae^{\alpha}+Be^{-2\alpha}
$$

那么

$$
\begin{aligned}
f(\alpha)&=Ae^{\alpha}+Be^{-2\alpha}\\
&=\frac{A}{2}e^{\alpha}+\frac{A}{2}e^{\alpha}+Be^{-2\alpha}\\
&\geq 3\sqrt[3]{\frac{A}{2}e^{\alpha}\times\frac{A}{2}e^{\alpha}\times Be^{-2\alpha}}\\
&=3\sqrt[3]{\frac{A^2B}{4}}
\end{aligned}
$$

当且仅当

$$
\frac{A}{2}e^{\alpha}=\frac{A}{2}e^{\alpha}=Be^{-2\alpha}
$$

时取等号，解得

$$
\alpha=\frac{1}{3}\ln\frac{2B}{A}
$$

(7) (vector calculus)

首先将式子展开：

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

注意A是对称矩阵，所以

$$
\begin{aligned}
\frac{\partial E(w)}{\partial w_k} &= \frac{1}{2}\sum_{j=1}^{d} A_{kj}w_j + \frac{1}{2}\sum_{i=1}^{d} w_i A_{ik} + b_k \\
&= \sum_{j=1}^{d} A_{kj}w_j + b_k \\
\frac{\partial^2 E(w)}{\partial w_l \partial w_k} &= \frac{\partial}{\partial w_l}\Big(\sum_{j=1}^{d} A_{kj}w_j + b_k\Big) \\
&= A_{kl} \\
&= A_{lk}
\end{aligned}
$$

写成矩阵形式，即得到

$$
\begin{aligned}
\nabla E(w) &= Aw + b \\
\nabla E^2(w) &= A
\end{aligned}
$$

(8) (quadratic programming)

令

$$\nabla E(w) = Aw + b = 0$$

可得

$$w = -A^{-1}b$$

又因为

$$\nabla E^2(w) = A$$

正定，所以在$w=-A^{-1}b$处取极小值。

(9) (optimization with linear constraint)

构造拉格朗日乘子：

$$L(w_1, w_2, w_3, \lambda) = \frac{1}{2}(w_1^2 + 2w_2^2 + 3w_3^2) - \lambda(w_1 + w_2 + w_3 - 1)$$

求偏导并令偏导数为0，可得

$$
\begin{aligned}
\frac{\partial L}{\partial w_1} &= w_1 - \lambda = 0 \\
\frac{\partial L}{\partial w_2} &= 2w_2 - \lambda = 0 \\
\frac{\partial L}{\partial w_3} &= 3w_3 - \lambda = 0 \\
\frac{\partial L}{\partial \lambda} &= -(w_1 + w_2 + w_3 - 1) = 0
\end{aligned}
$$

解得

$$
\begin{aligned}
w_1 &= \lambda \\
w_2 &= \frac{\lambda}{2} \\
w_3 &= \frac{\lambda}{3} \\
w_1 + w_2 + w_3 &= \frac{11}{6}\lambda = 1 \\
\lambda &= \frac{6}{11}
\end{aligned}
$$

带入可得

$$
\begin{aligned}
\min \frac{1}{2}(w_1^2 + 2w_2^2 + 3w_3^2) &= \frac{1}{2}\lambda^2(1 + \frac{1}{2} + \frac{1}{3}) \\
&= \frac{1}{2} \times \frac{6^2}{11^2} \times \frac{11}{6} \\
&= \frac{3}{11}
\end{aligned}
$$

(10) (optimization with linear constraints)

构造拉格朗日乘子：

$$
L(w, \lambda) = E(w) + \lambda^T(Aw + b)
$$

关于w求梯度并令其为0可得

$$
\nabla_w E(w) + \lambda^T A = 0
$$

所以结论成立。