Homework #0

DUE DATE: NONE

把n个元素分成两组：第一组n-1个,第二组1个
从中取出k个元素【方法有 C(n,k)种】，取法有两种
(1)从第一组中取出k个，方法有C(n-1,k)种
(2) 从第一组中取出k-1个,从第二组中取出1个，方法有C(n-1,k-1)种

1 Probability and Statistics

(1) (combinatorics)
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 $0 \leq K \leq N$.

(2) (counting)
What is the probability of getting exactly 6 heads when flipping 10 fair coins?

C(10,6)*P(Head)^4*P(Back)^6
=210*(0.5)^4*(0.5)^6

What is the probability of getting a full house (XXXYY) when randomly drawing 5 cards out of a deck of 52 cards?

There are C(52, 5) ways to choose 5 cards from 52. All these ways are equally likely. Now we will count the number of "full house" hands. For a full house, there are C(13, 1) ways to choose the kind we have three of. For each of these ways, the actual cards can be chosen in C(4, 3) ways. For each way of getting so far, there are C(12, 1) ways to choose the kind we have two of, and for each there are C(4, 2) ways to choose the actual cards. So our probability is C(13, 1)*C(4, 3)*C(12, 1)*C(4, 2)/C(52, 5)

(3) (conditional probabilit
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?

at least one is Head: P(all Head)/P(at least one Head)
(0.5)^3/(1-0.5^3)=1/7

(4) (Bayes theorem) P(B|A) = P(A|B)*P(B) / P(A) —— P(B|A)*P(A) = P(A|B)*P(B)
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 $\{0, 1, \ldots, 7\}$; 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?

(5) (union/intersection)
If $P(A) = 0.3$ and $P(B) = 0.4$,
what is the maximum possible value of $P(A \cap B)$?
what is the minimum possible value of $P(A \cap B)$?
what is the maximum possible value of $P(A \cup B)$?
what is the minimum possible value of $P(A \cup B)$?

0.3
0
0.7
0.4

D(X)=E{[X-E(X)]^2}=E(X^2) - [E(X)]^2
第一个式子：将第二个式子的右边展开,E[X-E(X)]^2=E[X^2-2XE(X)+(E(X))^2]=E(X^2)-2E(X)E(X)+(E(X))^2=E(X^2)-(E(X))^2

(6) (mean/variance)

Let mean $\overline{X} = \frac{1}{N}\sum_{n=1}^{N} X_n$ and variance $\sigma_X^2 = \frac{1}{N-1}\sum_{n=1}^{N}(X_n - \overline{X})^2$. Prove that

$$\sigma_X^2 = \frac{N}{N-1}\left(\frac{1}{N}\sum_{n=1}^{N} X_n^2 - \overline{X}^2\right).$$

(7) (Gaussian distribution)
If X_1 and X_2 are independent random variables, where $p(X_1)$ is Gaussian with mean 2 and variance 1, $p(X_2)$ is Gaussian with mean -3 and variance 4. Let $Z = X_1 + X_2$. Prove $p(Z)$ is Gaussian, and determine its mean and variance.

高斯分布/正态分布：若随机变量X服从一个数学期望为μ、方差为σ^2的正态分布，记为N(μ，σ^2)。其概率密度函数为正态分布的期望值μ决定了其位置，其标准差σ决定了分布的幅度。当μ=0,σ=1时的正态分布是标准正态分布。
公式：https://baike.baidu.com/item/正态分布/829892?fr=aladdin&fromid=10145793&fromtitle=%E9%AB%98%E6%96%AF%E5%88%86%E5%B8%83
如A ~N(μ1，Δ1²)，B~N(μ2，Δ2²)，且A，B相互独立，那么A+B~N(u1+μ2, Δ1²+Δ2²)。
证明过程：https://en.wikipedia.org/wiki/Sum_of_normally_distributed_random_variables

2 Linear Algebra

(1) (rank)

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

一个矩阵A的列秩是A的线性独立的纵列的极大数。通常表示为r(A)，rk(A)或rank A。m × n矩阵的秩最大为m和n中的较小者，表示为 min(m,n)。初等变化，看看有多少不是全零的行数：第一行*-1；然后分别与第二行、第三行相加；然后第二行的结果/2再与第三行相加

(2) (inverse)

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

设A是数域上的一个n阶方阵，若在相同数域上存在另一个n阶矩阵B，使得：AB=BA=E。则我们称B是A的逆矩阵，而A则被称为可逆矩阵。求逆矩阵的初等变换法：将一n阶可逆矩阵A和n阶单位矩阵I写成一个nX2n的矩阵 对B施行初等行变换，即对A与I进行完全相同的若干初等行变换，目标是把A化为单位矩阵。当A化为单位矩阵I的同时，B的右一半矩阵同时化为了A。通过若干次初等行变换（“某行乘以一个数后加到另一行”、“某两行互换位置”、“某行乘以某一个数”，这三种以行做运算的方法）

0 2 4 1 0 0　　1 0 0 a11 a12 a13
2 4 2 0 1 0 ——0 1 0 a21 a22 a23
3 3 1 0 0 1　　0 0 1 a31 a32 a33

设 A 是n阶方阵，如果存在数m和非零n维列向量 x，使得 Ax=mx 成立，则称 m 是矩阵A的一个特征值（characteristic value)或本征值（eigenvalue)。非零n维列向量x称为矩阵A的属于（对应于）特征值m的特征向量或本征向量. https://zhidao.baidu.com/question/454557811.html?qbl=relate_question_0

(3) (eigenvalues/eigenvectors)

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

(4) (singular value decomposition)

https://zhuanlan.zhihu.com/p/29846048 奇异值分解

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

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

(5) (PD/PSD)

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

(a) For any real matrix Z, $\mathrm{Z}\mathrm{Z}^T$ is PSD.

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

(6) (inner product)

http://dec3.jlu.edu.cn/webcourse/t000022/teach/chapter5/5_1.htm

Consider $\mathbf{x} \in R^d$ and some $\mathbf{u} \in R^d$ with $\|\mathbf{u}\| = 1$.
What is the maximum value of $\mathbf{u}^T\mathbf{x}$? What $\mathbf{u}$ results in the maximum value?
What is the minimum value of $\mathbf{u}^T\mathbf{x}$? What $\mathbf{u}$ results in the minimum value?
What is the minimum value of $|\mathbf{u}^T\mathbf{x}|$? What $\mathbf{u}$ results in the minimum value?

||x|| 同向
-||x|| 反向
0 正交

(7) (distance)

Consider two parallel hyperplanes in R^d:

$$H_1 : \mathbf{w}^T\mathbf{x} = +3,$$
$$H_2 : \mathbf{w}^T\mathbf{x} = -2,$$

where $\mathbf{w}$ is the norm vector. What is the distance between H_1 and H_2?

3 Calculus

y'=1/[(1+e^((-2)*x)]*[(e^((-2)x)+1]'=1/[e^ ((-2)x)+1]* (-2) *e^ ((-2) x)=(-2)*e^((-2)x)/(1+e^((-2)*x)

(1) (differential and partial differential)

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 $\frac{\partial g(x,y)}{\partial y}$?

(2) (chain rule)

(f(g(x)))' =f'(g(x))g'(x)
(sinx)' = cosx (cosx)' = - sinx
(tanx)'=1/(cosx)^2=(secx)^2=1+(tanx)^2
-(cotx)'=1/(sinx)^2=(cscx)^2=1+(cotx)^2

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

(3) (integral) 积分

What is $\int_5^{10} \frac{2}{x-3} dx$?

https://baike.baidu.com/item/梯度/13014729?fr=aladdin
https://baike.baidu.com/item/黑塞矩阵/2248782

(4) (gradient and Hessian)

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

https://baike.baidu.com/item/泰勒公式/7681487?fr=aladdin
https://zhidao.baidu.com/question/339541335.html

(5) (Taylor's expansion)

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

(6) (optimization)

For some given $A > 0, B > 0$, solve

$$\min_\alpha Ae^\alpha + Be^{-2\alpha}.$$

(7) (vector calculus)
Let $\mathbf{w}$ be a vector in R^d and $E(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T \mathrm{A}\mathbf{w} + \mathbf{b}^T\mathbf{w}$ for some symmetric matrix A and vector $\mathbf{b}$. Prove that the gradient $\nabla E(\mathbf{w}) = \mathrm{A}\mathbf{w} + \mathbf{b}$ and the Hessian $\nabla^2 E(\mathbf{w}) = \mathrm{A}$.

(8) (quadratic programming)
Following the previous question, if A is not only symmetric but also positive definite (PD), prove that the solution of $\mathrm{argmin}_{\mathbf{w}} E(\mathbf{w})$ is $-\mathrm{A}^{-1}\mathbf{b}$.

(9) (optimization with linear constraint)
Consider

$$\min_{w_1,w_2,w_3} \frac{1}{2}(w_1^2 + 2w_2^2 + 3w_3^2) \text{ subject to } w_1 + w_2 + w_3 = 11.$$

Refresh your memory on "Lagrange multipliers" and show that the optimal solution must happen on $w_1 = \lambda$, $2w_2 = \lambda$, $3w_3 = \lambda$. Use the property to solve the problem.

(10) (optimization with linear constraints)
Let $\mathbf{w}$ be a vector in R^d and $E(\mathbf{w})$ be a convex differentiable function of $\mathbf{w}$. Prove that the optimal solution to

$$\min_{\mathbf{w}} E(\mathbf{w}) \text{ subject to } \mathrm{A}\mathbf{w} + \mathbf{b} = 0.$$

must happen at $\nabla E(\mathbf{w}) + \boldsymbol{\lambda}^T \mathrm{A} = \mathbf{0}$ for some vector $\boldsymbol{\lambda}$. (*Hint: If not, let* $\mathbf{u}$ *be the residual when projecting* $\nabla E(\mathbf{w})$ *to the span of the rows of* A. *Show that for some very small* η, $\mathbf{w} - \eta \cdot \mathbf{u}$ *is a feasible solution that improves* E.)