UC Berkeley
Department of Electrical Engineering and Computer Science
Department of Statistics

EECS 281A / STAT 241A Statistical Learning Theory

Solutions to Problem Set 1

Fall 2012

Issued: Tuesday, August 28, 2012 **Due:** Tuesday, September 4, 2012

Reading: This homework is purely undergraduate review on probability and linear algebra; relevant material is covered in recitation on Wednesday, August 29th. If you are not familiar with the concepts here, then you ***do not have the appropriate background for this course,*** and will find the course too demanding. In this case, it would be best to drop the class, which would allow someone on the wait list with the appropriate background to enroll. This homework must be done alone, without consulting any classmates or friends.

Problem 1.1

Consider the collection of vectors:

$$\begin{bmatrix}1 & 1 & 1 & x\end{bmatrix}, \quad \begin{bmatrix}1 & 1 & x & 1\end{bmatrix}, \quad \begin{bmatrix}1 & x & 1 & 1\end{bmatrix}, \quad \begin{bmatrix}x & 1 & 1 & 1\end{bmatrix}.$$

(a) For which real numbers x do these vectors *not* form a basis of $\mathbb{R}^4$?

Solution: These vectors form a basis of $\mathbb{R}^4$ if and only if the basis matrix

$$B(x) = \begin{bmatrix}x & 1 & 1 & 1 \\ 1 & x & 1 & 1 \\ 1 & 1 & x & 1 \\ 1 & 1 & 1 & x\end{bmatrix}$$

is nonsingular. Using expansion by minors, we can compute

$$\det B(x) = x^4 - 6x^2 + 8x - 3 = (x-1)^3(x+3).$$

Therefore, the four vectors above do not form a basis of $\mathbb{R}^4$ if $x = 1$ or $x = -3$.

(b) For each value of x from (a), what is the dimension of the subspace of $\mathbb{R}^4$ that they span?

Solution: If $x = 1$ then the four vectors above are the same, so they span a one-dimensional subspace. If $x = 3$, then the first three vectors are linearly independent and their sum is equal to the fourth one, so the four vectors above span a three-dimensional subspace.

Problem 1.2
For square matrices, prove the following properties of the matrix trace and determinant:

(a) $\text{trace}(A + B) = \text{trace}(A) + \text{trace}(B)$.

Solution:

$$\text{trace}(A+B) = \sum_{i=1}^{N}(A+B)_{ii} = \sum_{i=1}^{N} A_{ii} + \sum_{i=1}^{N} B_{ii} = \text{trace}(A) + \text{trace}(B).$$

(b) $\text{trace}(AB) = \text{trace}(BA)$.

Solution:

$$\begin{aligned}\text{trace}(AB) &= \sum_{i=1}^{N}(AB)_{ii} = \sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}B_{ji} = \sum_{j=1}^{N}\sum_{i=1}^{N} B_{ji}A_{ij} \\ &= \sum_{j=1}^{N}(BA)_{jj} = \text{trace}(BA).\end{aligned}$$

(c) $\det(AB) = \det(A)\det(B)$.

Solution:

We solve this problem by considering two cases:

(a) $det(A) = 0$
(b) $det(B) \neq 0$.

Case 1: $det(A) = 0$
It suffices to show that $det(AB) = 0$. Since $det(A) = 0$, A does not have full-rank (n). Thus,

$$rank(AB) \leq \min\{rank(A), rank(B)\} \leq rank(A) < n.$$

In other words, (AB) does not have full-rank and so $det(AB) = 0$.

Case 2: $det(A) \neq 0$
The assumption that $det(A) \neq 0$ is equivalent to the assumption that A is the finite product of elementary matrices:

$$A = E_1 \cdots E_f.$$

Thus, if we can show that the results hold elementary matrices E_i:

$$det(E_iB) = det(E_i)det(B),$$

we will have

$$\begin{aligned} det(AB) &= det(E_1E_2\cdots E_fB) = det(E_1)det(E_2\cdots E_fB) = \cdots \\ &= det(E_1)\cdots det(E_f)det(B) \\ &= det(E_1E_2)det(E_3)\cdots det(B) \\ &= det(A)det(B) \end{aligned}$$

I will now show that the result holds for elementary matrices E.

Recall the possible choices for the single row operation one performs on an identity matrix in order to obtain an elementary matrix:

1. Switching the position of two rows: This operation makes the determinant of a matrix change sign.
2. Multiplying a row by a constant c: This operation makes the determinant of a matrix M change to $cdet(M)$.
3. Setting row $r_i = r_i + c * r_j$ for some constant c and distinct row R_j: This operation does nothing to the determinant.

It clearly follows that for an elementary matrix $E_i = O_i(I)$, $E_iB = O_i(B)$, where $O_i(M)$ is elementary row operation i just listed ($i \in \{1, 2, 3\}$) applied to matrix M. In other words, whatever row operation was used to obtain E_i from I can be applied to B to obtain E_iB.

It follows that

$$det(E_iB) = det(O_i(B)) = \begin{cases} -det(B) & \text{if } i = 1 \\ c * det(B) & \text{if } i = 2 \\ det(B) & \text{if } i = 3 \end{cases} = \begin{cases} det(E_i)det(B) & \text{if } i = 1 \\ det(E_i)det(B) & \text{if } i = 2 \\ det(E_i)det(B) & \text{if } i = 3 \end{cases},$$

as desired.

(d) For a non-singular matrix, $\det(A^{-1}) = 1/\det(A)$.

Solution: We have

$$1 = \det(I) = \det(AA^{-1}) = \det(A)\det(A^{-1}),$$

where I is the identity matrix, so $\det(A^{-1}) = 1/\det(A)$.

Problem 1.3

Consider the $(k+m)$-dimensional matrix $M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$, where $A \in \mathbb{R}^{k \times k}$, $D \in \mathbb{R}^{m \times m}$, $B \in \mathbb{R}^{k \times m}$ and $C \in \mathbb{R}^{m \times k}$.

(a) If $B = 0$ and $C = 0$, show that $\det(M) = \det(A)\det(D)$.

Solution: Note that if $B = 0$ and $C = 0$, then the eigenvalues of M are the union of the eigenvalues of A and the eigenvalues of D. Recalling that the determinant of a matrix is the product of its eigenvalues gives us the result.

(b) If A is invertible, show that $\det(M) = \det(A)\ \det(D - CA^{-1}B)$.

Solution: We first prove the statement when $B = 0$. Recall the Leibniz formula for determinant:

$$\det M = \sum_{\sigma \in S_{k+m}} \operatorname{sgn}(\sigma) \prod_{i=1}^{k+m} M_{i,\sigma_i}. \tag{1}$$

Here, S_{k+m} is the symmetric group of all permutations of the set $\{1, \dots, k+m\}$, and $\operatorname{sgn}(\sigma)$ is the sign of permutation σ (i.e. $\operatorname{sgn}(\sigma) = (-1)^\tau$ where τ is the number of transpositions in the decomposition of σ). Note that since $B = 0$, we have $M_{i,\sigma_i} = 0$ whenever $1 \leq i \leq k$ and $k+1 \leq \sigma_i \leq k+m$. Thus, the only nonzero terms in (1) are from the permutations σ that permute $\{1, \dots, k\}$ and $\{k+1, \dots, k+m\}$ separately. The set of such permutations is isomorphic to the product $S_k \times S_m$, and the signature of such permutations $\sigma = (\sigma^{(1)}, \sigma^{(2)}) \in S_k \times S_m$ can also be factorized as $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\sigma^{(1)})\ \operatorname{sgn}(\sigma^{(2)})$. Therefore,

$$\begin{aligned}
\det(M) &= \sum_{\sigma = (\sigma^{(1)}, \sigma^{(2)}) \in S_k \times S_m} \operatorname{sgn}(\sigma) \prod_{i=1}^{k+m} M_{i,\sigma_i} \\
&= \sum_{\sigma^{(1)} \in S_k} \sum_{\sigma^{(2)} \in S_m} \operatorname{sgn}(\sigma^{(1)})\ \operatorname{sgn}(\sigma^{(2)}) \prod_{i=1}^{k} M_{i,\sigma^{(1)}_i} \prod_{i=1}^{m} M_{k+i,\sigma^{(2)}_i} \\
&= \Big(\sum_{\sigma \in S_k} \operatorname{sgn}(\sigma) \prod_{i=1}^{k} M_{i,\sigma_i} \Big) \Big(\sum_{\sigma \in S_m} \operatorname{sgn}(\sigma) \prod_{i=1}^{m} M_{k+i,\sigma_i} \Big) \\
&= \det(A)\det(D).
\end{aligned}$$

The same method also works to prove the statement when $C = 0$.

Now noting that we can factorize

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & 0 \\ C & I \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ 0 & D - CA^{-1}B \end{bmatrix},$$

the general result follows from the multiplicativity of determinant and the two special cases above.

Problem 1.4
Prove or disprove: a symmetric matrix A is positive semidefinite if and only if $\text{trace}(A\,B) \geq 0$ for all symmetric positive semidefinite matrices B.
Solution: We prove the statement above. If $A \succeq 0$ (i.e. A is positive semidefinite), then for any $B \succeq 0$ we have

$$\text{trace}(AB) = \text{trace}(AB^{1/2}B^{1/2}) = \text{trace}(B^{1/2}AB^{1/2}) \geq 0,$$

since $B^{1/2}AB^{1/2}$ is also positive semidefinite. Conversely, suppose $\text{trace}(AB) \geq 0$ for all $B \succeq 0$. Taking $B = xx^\top$ for $x \in \mathbb{R}^n$, where n is the size of A, we get

$$x^\top Ax = \text{trace}(x^\top Ax) = \text{trace}(Axx^\top) = \text{trace}(AB) \geq 0.$$

This shows that $A \succeq 0$.

Problem 1.5
Let $X = (X_1, \ldots, X_n)$ be a jointly Gaussian random vector with mean $\mu \in \mathbb{R}^n$ and covariance matrix Σ. Let $W = (W_1, \ldots, W_n)$ be a second jointly Gaussian random vector with mean ν and covariance Λ.

(a) Let A and B be $n \times n$ matrices, and form the random vector $Y = AX + BW$. Compute the mean vector and covariance matrix of Y. (Your answer can involve $\text{cov}(X, W)$.)

Solution: The mean vector of Y is

$$\mathbb{E}[Y] = A\mathbb{E}[X] + B\mathbb{E}[W] = A\mu + B\nu,$$

and the covariance matrix is

$$\begin{aligned} \text{cov}(Y) &= \text{cov}(AX + BW, AX + BW) \\ &= A\,\text{cov}(X, X)A^\top + A\,\text{cov}(X, W)B^\top + B\,\text{cov}(W, X)A^\top + B\,\text{cov}(W, W)B^\top \\ &= A\Sigma A^\top + A\,\text{cov}(X, W)B^\top + B\,\text{cov}(X, W)^\top A^\top + B\Lambda B^\top. \end{aligned}$$

(b) How does your answer change if X and W are uncorrelated?

Solution: If $\text{cov}(X, W) = 0$, then $\mathbb{E}[Y]$ is the same as in part (a) above, but the covariance matrix now becomes

$$\text{cov}(Y) = A\Sigma A^\top + B\Lambda B^\top.$$

Problem 1.6

Craig is doing a study of moose in the Alaskan wilderness, and wants to estimate their heights. Let X be the height in meters of a randomly selected moose. Craig is interested in estimating $h = \mathbb{E}[X]$. Being sure that no moose is taller than 3 meters, Craig decides to use 1.5 meters as a conservative (large) value for the standard deviation of X. To estimate h, Craig compute the average H of the heights of n moose that he selects at random.

(a) Compute $\mathbb{E}[H]$ and $\text{var}(H)$ in terms of h and Craig's 1.5 meter bound for $\text{std}(X)$.

Solution: Let $X_1, \dots, X_n$ be the heights of the moose that Craig selects at random, so $X_1, \dots, X_n$ are i.i.d. with the same distribution as X, and we can write $H = \frac{1}{n}\sum_{i=1}^{N} X_i$. The expectation of H is

$$\mathbb{E}[H] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mathbb{E}[X] = h,$$

and, assuming $\text{std}(X) = 1.5$, the variance of H is

$$\text{var}(H) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i) = \frac{\text{var}(X)}{n} = \frac{1.5^2}{n}.$$

(b) Compute the minimum value of n (with $n > 0$) such that the standard deviation of H will be less than 0.01 meters.

Solution: We want

$$\text{std}(H) = \frac{1.5}{\sqrt{n}} < 0.01 \quad \Longleftrightarrow \quad n > 150^2,$$

so the minimum such n is $1 + 150^2 = 22,501$.

(c) Say Craig would like to be 99% sure that his estimate is within 5 centimeters of the true average height of moose. Using the Chebyshev inequality, calculate the minimum value of n required.

Solution: By Chebyshev inequality,

$$\mathbb{P}(|H-h| \geq 0.05) \leq \frac{\operatorname{var}(H)}{0.05^2} = \frac{900}{n},$$

so for H to be within 5 centimeters of h with 99% certainty, Craig would need

$$1 - \frac{900}{n} \geq 0.99 \quad \Longleftrightarrow \quad n \geq 90,000.$$

(d) If we agree that no moose are taller than three meters, why is it correct to use 1.5 meters as an upper bound on the standard deviation for X, the height of any moose selected at random?

Solution: Recall that $h = \mathbb{E}[X]$ is the value that minimizes the function $x \mapsto \mathbb{E}[(X-x)^2]$. Therefore, since $0 \leq X \leq 3$,

$$\operatorname{var}(X) = \mathbb{E}[(X-h)^2] \leq \mathbb{E}\Big[\Big(X - \frac{3}{2}\Big)^2\Big] \leq \Big(\frac{3}{2}\Big)^2 = \frac{9}{4},$$

and hence $\operatorname{std}(X) \leq 3/2 = 1.5$.

Problem 1.7

A group of N archers shoot at a target. The distance of each shot from the center of the target is uniformly distributed between 0 to 1, independently of the other shots.

(a) Find the expected distance from the winner's arrow to the center. (The winner's arrow is closest to the origin.)

Solution: Let $U_1, \ldots, U_N \overset{\text{i.i.d.}}{\sim} \text{Uniform}([0,1])$ denote the distances of the shots from the center of the target, and let $W_N = \min(U_1, \ldots, U_N)$ denote the distance from the winner's arrow to the center. Then

$$\begin{aligned}\mathbb{E}[W_N] &= \int_0^1 \mathbb{P}(W_N \geq t)\, dt = \int_0^1 \prod_{i=1}^N \mathbb{P}(U_i \geq t)\, dt \\ &= \int_0^1 (1-t)^N\, dt = \frac{1}{N+1}.\end{aligned}$$

(b) Find the expected distance from the loser's arrow to the center. (The loser's arrow is the arrow farthest away from the origin).

Solution: Let $L_N = \max(U_1, \ldots, U_N)$ denote the distance from the loser's arrow to the center. Clearly

$$\mathbb{P}(L_N \le t) = \prod_{i=1}^{N} \mathbb{P}(U_i \le t) = t^N \quad \text{for } 0 \le t \le 1,$$

so L_N has density function $f_{L_N}(t) = Nt^{N-1}$. Therefore,

$$\mathbb{E}[L_N] = \int_0^1 t f_{L_N}(t)\, dt = \int_0^1 Nt^N\, dt = \frac{N}{N+1}.$$

Problem 1.8

Every day that he leaves work, Fred the Absent-minded Accountant toggles his light switch according to the following protocol: (i) if the light is on, he switches it off with probability 0.60; and (ii) if the light is off, he switches it on with probability 0.20. At no other time (other than the end of each day) is the light switch touched.

(a) Suppose that on Monday night after leaving work, Fred's office is equally likely to be light or dark. What is the probability that his office will be lit all five nights of the week (Monday through Friday)?

Solution:

Number Monday - Friday consecutively from 1 - 5. Let $X_i = 1$ if the light is on the evening of day i and 0 otherwise. It is given that the probability the light is left on Monday night is 1/2. Hence

$$\begin{aligned} P[X_1, \ldots, X_5 = 1] &= P[X_1 = 1]P[X_2, \ldots, X_5 = 1] \\ &= P[X_1 = 1]P[X_2 = 1]^4 \\ &= (1/2)(1 - .6)^4 = .0128 \end{aligned}$$

(b) Suppose that you observe that his office is lit on both Monday and Friday nights after work. Compute the expected number of nights, from that Monday through Friday, that his office is lit.

Solution:

Number Monday - Friday consecutively from 1 - 5. Let $X_i = 1$ if the light is on the evening of day i and 0 otherwise.

$$P(X_2, X_3, X_4 \mid X_1 = 1, X_5 = 1) = \frac{P(X_2, X_3, X_4, X_5 = 1 \mid X_1 = 1)}{P(X_5 = 1 \mid X_1 = 1)} \tag{2}$$

Also,

$$\begin{aligned}
E[\textstyle\sum_{i=1}^n X_i] &= 2P(X_2=0,X_3=0,X_4=0 \mid X_1=1,X_5=1) \\
&\quad +3[P(X_2=1,X_3=0,X_4=0 \mid X_1=1,X_5=1)+ \\
&\qquad P(X_2=0,X_3=1,X_4=0 \mid X_1=1,X_5=1)+ \\
&\qquad P(X_2=0,X_3=0,X_4=1 \mid X_1=1,X_5=1)] \\
&\quad +4[P(X_2=1,X_3=1,X_4=0 \mid X_1=1,X_5=1)+ \\
&\qquad P(X_2=1,X_3=0,X_4=1 \mid X_1=1,X_5=1)+ \\
&\qquad P(X_2=0,X_3=1,X_4=1 \mid X_1=1,X_5=1)] \\
&\quad +5P(X_2=1,X_3=1,X_4=1 \mid X_1=1,X_5=1)
\end{aligned}$$

Let $z = P(X_5 = 1 \mid X_1 = 1)$, then applying 2 we get that, for

$$\begin{aligned}
zE[\textstyle\sum_{i=1}^n X_i] &= 2P(X_2=0,X_3=0,X_4=0,X_5=1 \mid X_1=1) \\
&\quad +3[P(X_2=1,X_3=0,X_4=0,X_5=1 \mid X_1=1)+ \\
&\qquad P(X_2=0,X_3=1,X_4=0,X_5=1 \mid X_1=1)+ \\
&\qquad P(X_2=0,X_3=0,X_4=1,X_5=1 \mid X_1=1)] \\
&\quad +4[P(X_2=1,X_3=1,X_4=0,X_5=1 \mid X_1=1)+ \\
&\qquad P(X_2=1,X_3=0,X_4=1,X_5=1 \mid X_1=1)+ \\
&\qquad P(X_2=0,X_3=1,X_4=1,X_5=1 \mid X_1=1)] \\
&\quad +5P(X_2=1,X_3=1,X_4=1,X_5=1 \mid X_1=1) \\
&= 2(.6)(.8)(.8)(.2)+3\,[(.4)(.6)(.8)(.2)+(.6)(.2)(.8)(.2)+(.6)(.8)(.2)(.4)] \\
&\quad +4\,[(.4)(.4)(.6)(.2)+(.4)(.6)(.2)(.4)+(.6)(.2)(.4)(.4)]+5(.4)(.4)(.4)(.4) \\
&= .8
\end{aligned}$$

Solving for z:

$$\begin{aligned}
z &= P(X_2=0,X_3=0,X_4=0,X_5=1 \mid X_1=1) \\
&\quad +[P(X_2=1,X_3=0,X_4=0,X_5=1 \mid X_1=1)+ \\
&\qquad P(X_2=0,X_3=1,X_4=0,X_5=1 \mid X_1=1)+ \\
&\qquad P(X_2=0,X_3=0,X_4=1,X_5=1 \mid X_1=1)] \\
&\quad +[P(X_2=1,X_3=1,X_4=0,X_5=1 \mid X_1=1)+ \\
&\qquad P(X_2=1,X_3=0,X_4=1,X_5=1 \mid X_1=1)+ \\
&\qquad P(X_2=0,X_3=1,X_4=1,X_5=1 \mid X_1=1)] \\
&\quad +P(X_2=1,X_3=1,X_4=1,X_5=1 \mid X_1=1) \\
&= (.6)(.8)(.8)(.2)+[(.4)(.6)(.8)(.2)+(.6)(.2)(.8)(.2)+(.6)(.8)(.2)(.4)] \\
&\quad +[(.4)(.4)(.6)(.2)+(.4)(.6)(.2)(.4)+(.6)(.2)(.4)(.4)]+(.4)(.4)(.4)(.4) \\
&= .256
\end{aligned}$$

Hence, the expected number of nights the light will be on is

$$E\left[\sum_{i=1}^n X_i\right] = .8/.256 = 3.125.$$

(c) Suppose that Fred's office is lit on Monday night after work. Compute the expected number of days until the first night that his office is dark.

Solution: Let L denotes the number of days until the first night his office is dark. Then

$$\begin{aligned}
\mathbb{P}(L=1) &= \mathbb{P}(x_2=0|x_1=1)=0.6 \\
\mathbb{P}(L=2) &= \mathbb{P}(x_3=0,x_2=1|x_1=1)=0.4\times 0.6 \\
\mathbb{P}(L=3) &= \mathbb{P}(x_4=0,x_3=1,x_2=1|x_1=1)=0.4\times 0.4^2\times 0.6 \\
\mathbb{P}(L=4) &= \mathbb{P}(x_5=0,x_4=1,x_3=1,x_2=1|x_1=1)=0.4^3\times 0.6 \\
\mathbb{P}(L=5) &= \mathbb{P}(x_5=1,x_4=1,x_3=1,x_2=1|x_1=1)=0.4^4
\end{aligned}$$

Thus, the expectation is

$$\begin{aligned}
EL &= \sum_{l=1}^{5} l\cdot\mathbb{P}(L=l) \\
&= 1\times 0.6+2\times 0.4\times 0.6+3\times 0.4^2\times 0.6+4\times 0.4^3\times 0.6+5\times 0.4^4 \\
&= 1.6496
\end{aligned}$$

Now suppose that Fred has been working for five years (i.e., assume that the Markov chain is in steady state).

(d) Is his light more likely to be on or off at the end of a given workday?

Solution: We have transition matrix

$$P=\begin{bmatrix} .8 & .2 \\ .6 & .4 \end{bmatrix}.$$

We wish to find the stable distribution (π_0,π_1), where .

$$\begin{aligned}
\pi_0 &= \mathbb{P}(x=0) \\
\pi_1 &= \mathbb{P}(x=1)
\end{aligned}$$

We know that

$$(\pi_0,\pi_1)P=(\pi_0,\pi_1).$$

And we know that $\pi_0 + \pi_1 = 1$ so we can get:

$$\pi_0 = 0.75$$
$$\pi_1 = 0.25$$

so the light is more likely to be off.