Homework #0 - A

Spring 2020, CSE 446/546: Machine Learning
Richy Yun
Due: 4/8/19 11:59 PM

Probability and Statistics

A.1 *[2 points]* According to the problem, we have:

$$p(x = 1|y = 1) = 0.99 \qquad p(x == 1|y == 0) = 0.01 \qquad p(y = 1) = 0.0001 \qquad p(y = 1) = 0.9999$$

where $x = 1$ is the event the test is positive, $y = 1$ is the event you have the disease, and $y = 1$ is the event you do not have the disease. Using Bayes rule, the probability of having the disease when you test positive is:

$$\begin{aligned} p(y = 1|x = 1) &= \frac{p(x = 1|y = 1)p(y = 1)}{p(x = 1|y = 1)p(y = 1) + p(x = 1|y = 0)p(y = 0)} \\ &= \frac{0.99 \times 0.001}{0.99 \times 0.0001 + 0.01 \times 0.9999} = 0.0098 \end{aligned}$$

A.2

a. *[1 points]* We can rewrite the covariance as:

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - Y\mathbb{E}[X] - X\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[Y]\mathbb{E}[X] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

if $\mathbb{E}[Y|X = x] = x$ then using the law of total expectation:

$$\begin{aligned} \mathbb{E}[Y|X = x] &= x \\ \mathbb{E}[\mathbb{E}[Y|X = x]] &= \mathbb{E}[x] \\ \mathbb{E}[Y] &= \mathbb{E}[x] = \mathbb{E}[X] \end{aligned}$$

We can also rewrite $\mathbb{E}[XY]$ as:

$$\begin{aligned} \mathbb{E}[XY] &= \sum_x \sum_y xyp(x, y) \\ p(x, y) &= p(y|x)p(x) \\ \mathbb{E}[XY] &= \sum_x \sum_y xyp(y|x)p(x) \\ &= \sum_x xp(x) \sum_y yp(y|x) \\ &= \sum_x xp(x)\mathbb{E}(Y|X = x) \\ &= \sum_x x^2 p(x) \\ &= \mathbb{E}[X^2] \end{aligned}$$

Thus, we now have:

$$\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \text{Cov}(X,X) = \mathbb{E}[(X-\mathbb{E}[X])^2]$$

b. *[1 points]* From (a):

$$\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Thus,

$$\text{Cov}(X,Y) = 0$$

A.3

a. *[2 points]* To calculate the PDF you can first find the CDF and take the derivative:

$$CDF_Z(z) = p[X+Y \leq z]$$

where p is probability. Using the rule of total probability:

$$\begin{aligned} CDF_Z(z) &= \int_{-\infty}^{\infty} p[X+Y \leq z | X = x] PDF_X(x) dx \\ &= \int_{-\infty}^{\infty} p[x+Y <= z] f(x) dx \end{aligned}$$

But using the definition of CDF we can now see:

$$p[x+Y \leq z] = p[Y \leq z-x] = CDF_Y(z-x)$$

Now we can take the derivative to find the PDF:

$$\begin{aligned} \frac{d}{dz} CDF_Z(z) = PDF_Z(z) = h(z) &= \frac{d}{dz} \int_{-\infty}^{\infty} CDF_Y(z-x) f(x) dx \\ &= \int_{-\infty}^{\infty} \frac{d}{dz} CDF_Y(z-x) f(x) dx \\ &= \int_{-\infty}^{\infty} PDF_Y(z-x) f(x) dx \\ &= \int_{-\infty}^{\infty} g(z-x) f(x) dx \end{aligned}$$

b. *[1 points]* The PDF of the sum of two random variables is a convolution as shown in part (a). Thus, we simply need to convolve the two uniform distributions together:

$$\begin{aligned} h(z) &= \int_{-\infty}^{\infty} g(z-x) f(x) dx \\ &= \begin{cases} z+1 & -1 \leq z < 0 \\ 1-z & 0 \leq z < 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

A.4 *[1 points]* $\mathcal{N}(0,1)$ indicates mean of 0 and variance of 1. With linear transformations of gaussian distributed random variables we know (from lecture and Murphy):

$$\text{If} \qquad Y = aX + b, \qquad \mu_Y = a\mu_X + b \qquad \text{and} \qquad \text{var}_Y = a^2 \text{var}_X$$

where Y and X are gaussian distributed random variables and a and b are constants. Since only a affects the variance, we can first find a such that $\text{var}_X = 1$:

$$1 = a^2\sigma^2 \qquad \rightarrow \qquad a = \frac{1}{\sigma}$$

Next, b must offset the newly scaled mean, and thus:

$$0 = \frac{1}{\sigma}\mu + b \qquad \rightarrow \qquad b = -\frac{1}{\sigma}\mu$$

A.5 *[2 points]* If X and Y are independent and identically distributed, then we know (again from lecture and Murphy):

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y] \qquad \text{and} \qquad \text{var}[X+Y] = \text{var}[X] + \text{var}[Y]$$

Using these in conjunction with how the mean and variance change due to a linear transformation (as in A.4) we know:

$$\text{mean}[\hat{\mu_n}] = \mu \qquad \text{and} \qquad \text{variance}[\hat{\mu_n}] = \frac{1}{n}\sigma^2$$

Applying another linear transformation again changes the mean and variance, in this case $a = \sqrt{n}$ and $b = -\mu\sqrt{n}$:

$$\text{mean} = a\mu + b = 0 \qquad \text{and} \qquad \text{variance} = a^2\frac{1}{n}\sigma^2 = \sigma^2$$

A.6

a. *[1 points]* The expected value of a sum of independent variables is the sum of the expected value of each random variable:

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Thus,

$$\mathbb{E}[\hat{F_n}(x))] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{X_i \leq x\}\right] = \frac{1}{n}(\mathbb{E}[\mathbf{1}\{X_1 \leq x\}] + \cdots + \mathbb{E}[\mathbf{1}\{X_n \leq x\}]) = F(x)$$

b. *[1 points]* To find the variance of $\hat{F_n}(x)$, we can first find the variance of $\mathbf{1}\{X_i \leq x\}$. Since it is a Bernoulli distribution, we know (from Murphy page 34):

$$\text{variance}[\mathbf{1}\{X_i \leq x\}] = \mathbb{E}[\mathbf{1}\{X_i \leq x\}](1 - \mathbb{E}[\mathbf{1}\{X_i \leq x\}]) = F(x)(1 - F(x))$$

We can see that this holds true by writing out the variance:

$$\begin{aligned}
\text{variance}[\mathbf{1}\{X \leq x\}] &= \mathbb{E}[(\mathbf{1}\{X \leq x\})^2] - \mathbb{E}[\mathbf{1}\{X \leq x\}]^2 \\
&= \mathbf{1}\{X \leq x | X \leq x\}^2 p(X \leq x) + \mathbf{1}\{X > x | X > x\}^2 p(X > x) - F(x)^2 \\
&= 1 \times p(X \leq x) + 0 \times p(X > x) - F(x)^2 \\
&= \mathbf{1}\{X \leq x | X \leq x\} p(X \leq x) - F(x)^2 \\
&= F(x) - F(x)^2 = F(x)(1 - F(x))
\end{aligned}$$

Now, knowing how the variance changes due to a linear transformation, we can show:

$$\text{variance}[\hat{F_n}(x)] = \frac{1}{n^2} nF(x)(1 - F(x)) = \frac{F(x)(1 - F(x))}{n}$$

c. *[1 points]* We first rearrange the expression:

$$\begin{aligned}
\frac{F(x)(1 - F(x))}{n} &\leq \frac{1}{4n} \\
F(x)(1 - F(x)) &\leq \frac{1}{4} \\
4F(x)^2 - 4F(x) + 1 &\geq 0 \\
(2F(x) - 1)^2 &\geq 0
\end{aligned}$$

Since squaring a value is always positive, the above must hold true, meaning the initial expression holds true as well.

Linear Algebra and Vector Calculus

A.7

a. *[2 points]* The rank of both matrices is 2. For matrix A, the third column is three times the first column minus the second column, so there are only two linearly independent columns. For matrix B, the third column is the sum of the first two columns so again there are only two linearly independent columns.

b. *[2 points]* The basis for the column span for both matrices is the first two columns since, as mentioned above, the third column is a linear combination of the two:

$$\text{basis} = \begin{bmatrix} 1 & 2 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

A.8

a. *[1 points]* This problem requires basic matrix multiplication. The row/column of the resulting matrix is the dot product of the row from the first matrix and the column from the second.

$$\begin{bmatrix} 0 & 2 & 4 \\ 2 & 4 & 2 \\ 3 & 3 & 1 \end{bmatrix} * \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} (0,2,4)\cdot(1,1,1) \\ (2,4,2)\cdot(1,1,1) \\ (3,3,1)\cdot(1,1,1) \end{bmatrix} = \begin{bmatrix} 6 \\ 8 \\ 7 \end{bmatrix}$$

b. *[2 points]* To solve, we need to find the inverse of matrix A:

$$Ax = b \qquad \rightarrow \qquad x = A^{-1}b$$

First, we find the determinant both to make sure it's not 0 so it can be inverted and for use later:

$$\det(A) = 0(4-6) - 2(2-6) + 4(6-12) = -16$$

Next, we transpose the matrix and find the cofactors (determinants of each minor matrix with correct signs applied):

$$A^T = \begin{bmatrix} 0 & 2 & 3 \\ 2 & 4 & 3 \\ 4 & 2 & 1 \end{bmatrix}$$

$$\text{cofactor}(A^T) = \begin{bmatrix} 4-6 & 2-12 & 4-16 \\ 2-6 & 0-12 & 0-8 \\ 6-12 & 0-6 & 0-4 \end{bmatrix} .* \begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix} = \begin{bmatrix} -2 & 10 & -12 \\ 4 & -12 & 8 \\ -6 & 6 & -4 \end{bmatrix}$$

where $.*$ represents element-wise multiplication similar to Matlab. Finally, we multiply the resulting matrix by one over the determinant:

$$A^{-1} = -\frac{1}{16}\begin{bmatrix} -2 & 10 & -12 \\ 4 & -12 & 8 \\ -6 & 6 & -4 \end{bmatrix} = \begin{bmatrix} -0.125 & -0.625 & 0.75 \\ -0.25 & 0.75 & -0.5 \\ 0.375 & -0.375 & 0.25 \end{bmatrix}$$

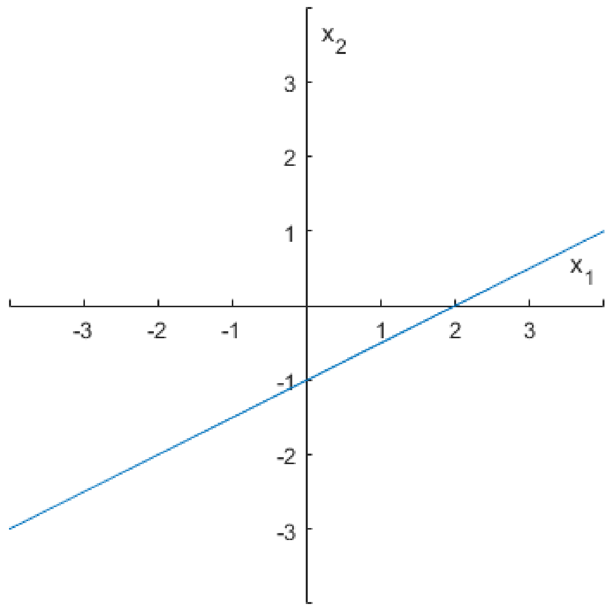
Now we multiply A^{-1} with b to obtain x:

$$x = \begin{bmatrix} -0.125 & -0.625 & 0.75 \\ -0.25 & 0.75 & -0.5 \\ 0.375 & -0.375 & 0.25 \end{bmatrix} * \begin{bmatrix} -2 \\ -2 \\ -4 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \\ -1 \end{bmatrix}$$

A.9 (Hyperplanes)

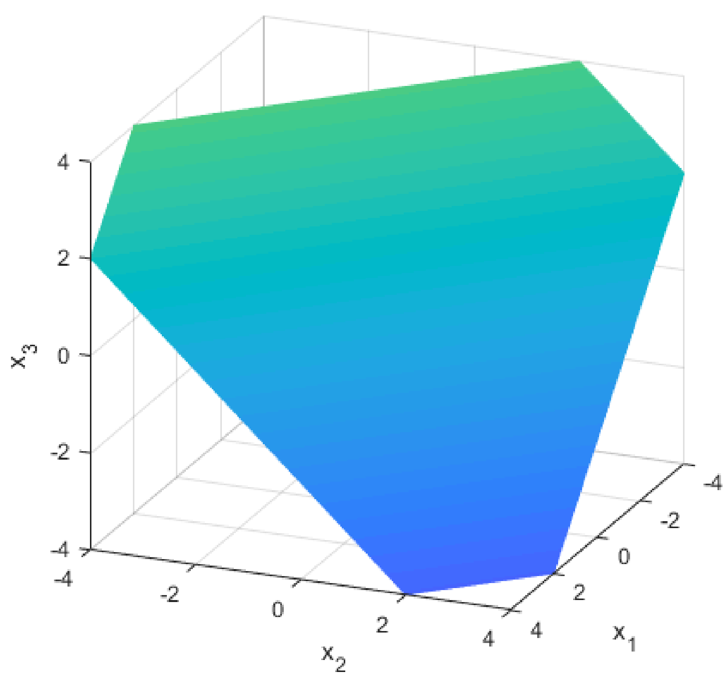
a. *[1 points]* ($n = 2$ example) Plugging in the variables gives us a line:

$$-x_1 + 2x_2 = -2 \qquad \rightarrow \qquad x_1 - 2x_2 - 2 = 0$$



b. *[1 points]* ($n = 3$ example) Plugging in the variables gives us a plane:

$$x_1 + x_2 + x_3 - 2 = 0$$



c. *[2 points]* Since $\widetilde{x}_0$ must fulfill $w^T x + b = 0$, we know $w^T \widetilde{x}_0 = -b$. Thus:

$$\|x_0 - \widetilde{x}_0\| = \left| \frac{w^T x_0 + b}{\|w\|} \right| \qquad \rightarrow \qquad \|x_0 - \widetilde{x}_0\|^2 = \frac{(w^T x_0 + b)^2}{\|w\|^2}$$

A.10

a. *[2 points]* Assuming x and y are vectors and c is a constant, we know x and y have to be vectors of size $1 \times n$ to ensure the matrix multiplication works. Therefore:

$$\begin{aligned}
(\boldsymbol{A}x)_i &= \sum_{k=1}^{n} A_{ik} x_k \\
x^T \boldsymbol{A} x &= \sum_{m=1}^{n} x_m \sum_{k=1}^{n} A_{mk} x_k = \sum_{m=1}^{n} \sum_{k=1}^{n} x_m A_{mk} x_k \\
f(x, y) &= \sum_{m=1}^{n} \sum_{k=1}^{n} x_m A_{mk} x_k + \sum_{m=1}^{n} \sum_{k=1}^{n} y_m B_{mk} x_k + c
\end{aligned}$$

b. *[2 points]* To approach this problem for the summation over indices, we can try thinking about what would happen element by element. For the ith partial derivative:

$$\nabla_x f(x, y)_i = \frac{\partial}{\partial x_i} \left(\sum_{m=1}^{n} \sum_{k=1}^{n} x_m A_{mk} x_k + \sum_{m=1}^{n} \sum_{k=1}^{n} y_m B_{mk} x_k + c \right)$$

We can apply the partial derivative to each part of the sum. The first term requires the product rule. Since all $x_i \neq x_1$ has a partial derivative of 0, we can simplify the sums to just the ith element. Therefore we can determine the summation over indices notation:

$$\begin{aligned}
\nabla_x f(x, y)_i &= \frac{\partial}{\partial x_i} \left(\sum_{m=1}^{n} \sum_{k=1}^{n} x_m A_{mk} x_k + \sum_{m=1}^{n} \sum_{k=1}^{n} y_m B_{mk} x_k + c \right) \\
&= \frac{\partial}{\partial x_i} \sum_{m=1}^{n} \sum_{k=1}^{n} x_m A_{mk} x_k + \frac{\partial}{\partial x_i} \sum_{m=1}^{n} \sum_{k=1}^{n} y_m B_{mk} x_k + \frac{\partial}{\partial x_i} c \\
&= \sum_{m=1}^{n} \sum_{k=1}^{n} \frac{\partial x_m}{\partial x_i} A_{mk} x_k + \sum_{m=1}^{n} \sum_{k=1}^{n} x_m A_{mk} \frac{\partial x_k}{\partial x_i} + \sum_{m=1}^{n} \sum_{k=1}^{n} y_m B_{mk} \frac{\partial x_k}{\partial x_i} \\
&= \sum_{k=1}^{n} A_{ik} x_k + \sum_{m=1}^{n} x_m A_{mi} + \sum_{m=1}^{n} y_m B_{mi}
\end{aligned}$$

We can convert this to vector notation by realizing that each sum is matrix multiplication as shown in part (a):

$$\begin{aligned}
\nabla_x f(x, y) &= \boldsymbol{A}x + x^T \boldsymbol{A} + y^T \boldsymbol{B} \\
&= \boldsymbol{A}x + \boldsymbol{A}^T x + y^T \boldsymbol{B} \\
&= \left(\boldsymbol{A} + \boldsymbol{A}^T \right) x + y^T \boldsymbol{B}
\end{aligned}$$

c. *[2 points]* We can approach this problem similar to part (b).

$$\begin{aligned}
\nabla_y f(x, y)_i &= \frac{\partial}{\partial y_i} \left(\sum_{m=1}^{n} \sum_{k=1}^{n} x_m A_{mk} x_k + \sum_{m=1}^{n} \sum_{k=1}^{n} y_m B_{mk} x_k + c \right) \\
&= \frac{\partial}{\partial y_i} \sum_{m=1}^{n} \sum_{k=1}^{n} x_m A_{mk} x_k + \frac{\partial}{\partial y_i} \sum_{m=1}^{n} \sum_{k=1}^{n} y_m B_{mk} x_k + \frac{\partial}{\partial y_i} c \\
&= \sum_{m=1}^{n} \sum_{k=1}^{n} \frac{\partial y_m}{\partial y_i} B_{mk} x_k \\
&= \sum_{k=1}^{n} B_{ik} x_k
\end{aligned}$$

And again we can convert this to vector notation using matrix multiplication:

$$\nabla_y f(x, y) = \boldsymbol{B}x$$

A.11

a. *[2 points]*

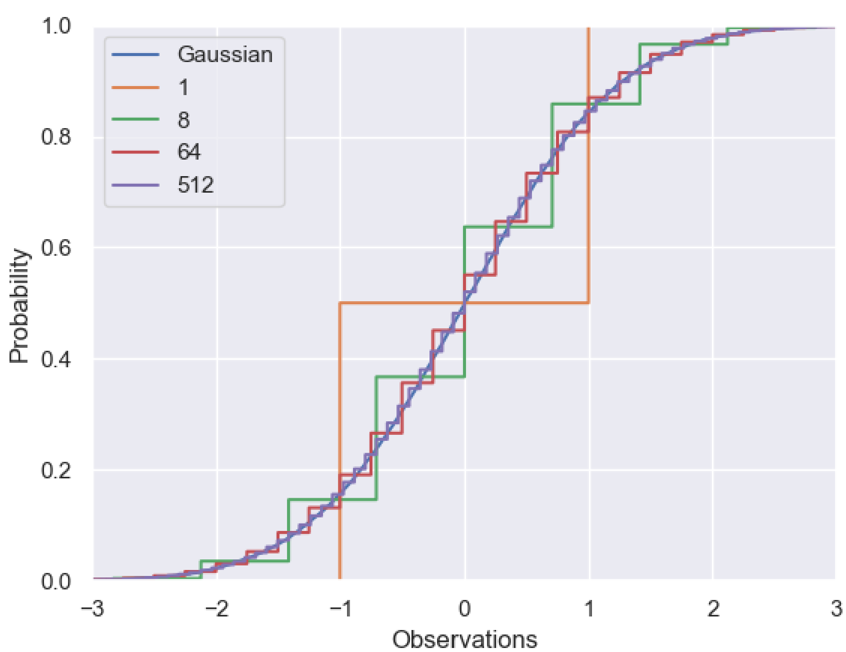
```
Ainv=
 [[ 0.125 -0.625  0.75 ]
 [-0.25   0.75  -0.5  ]
 [ 0.375 -0.375  0.25 ]]
```

b. *[1 points]*

```
Ainv*b=
 [[-2.]
 [ 1.]
 [-1.]]

A*c=
 [[6]
 [8]
 [7]]
```

A.12 *[4 points]* Since in problem A.6 we determined $\mathbb{E}[(\widehat{F}_n(x) - F(x))^2] \leq \frac{1}{4n}$, for $\sqrt{\mathbb{E}[(\widehat{F}_n(x) - F(x))^2]} \leq 0.0025$, $n = 40000$.



Source Code

A.9 (Matlab)

```
% A9 figures
% a)
x = linspace(-5,5,100);
y = (x-2)/2;
figure; plot(x,y)
set(gca,'xaxislocation','origin')
set(gca,'yaxislocation','origin')
box off
xlim([-4,4])
ylim([-4,4])
xlabel('x_1')
ylabel('x_2')

% b)
x = linspace(-5,5,100);
y = linspace(-5,5,100);
[X,Y] = meshgrid(x,y);
Z = 2-X-Y;
figure; surf(X,Y,Z);
shading interp
xlim([-4,4]); ylim([-4,4]); zlim([-4,4])
xlabel('x_1'); ylabel('x_2'); zlabel('x_3');
```

A.11 (Python)

```
import numpy as np

# A.11 (a)
A = np.matrix('0,2,4;2,4,2;3,3,1')
Ainv = np.linalg.inv(A)
print('\nAinv=\n',Ainv)

# A.11 (b)
b = np.matrix('-2;-2;-4')
c = np.matrix('1;1;1')
print('\nAinv*b=\n',np.matmul(Ainv,b))
print('\nA*c=\n',np.matmul(A,c))
```

A.12 (Python)

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()

n=40000 # From problem A.6

#(a)
Z=np.random.randn(n)
```

```
plt.step(sorted(Z), np.arange(1,n+1)/float(n))

#(b)
k = [1,8,64,512]
for x in k:
Y = np.sum(np.sign(np.random.randn(n, x))*np.sqrt(1./x), axis=1)
plt.step(sorted(Y), np.arange(1,n+1)/float(n))

plt.xlim(-3,3)
plt.ylim(0,1)
plt.xlabel("Observations")
plt.ylabel("Probability")
plt.legend(['Gaussian','1','8','64','512'])
plt.show()
```