OSU CSE 3521/5521
Homework #3: Problem Set

Release Date: October 30th, 2020

Submission Instructions

Due Date: November 16st (23:59 ET), 2020

Submission: Please submit your solutions in a single PDF file named HW_3_name.number.pdf (e.g., HW_3_chao.209.pdf) to Carmen. You may write your solutions on paper and scan it, or directly type your solutions and save them as a PDF file. *Submission in any other format will not be graded.*

We highly recommend that you write down the derivation of your answers, and highlight your answers clearly!

Collaboration: You may discuss with your classmates. However, you need to write your own solutions and submit them separately. Also in your written report, you need to list with whom you have discussed for each problem. Please consult the syllabus for what is and is not acceptable collaboration.

Calculation: Please perform rounding to your results after the second decimal number. For example, 1.245 becomes 1.25 and -1.228 becomes -1.23.

1 Regression, Gauss-Newton, and gradient descent: Part-1 [18 points]

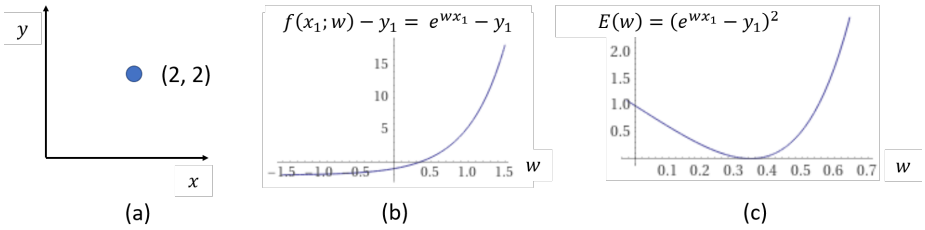


Figure 1: (a) A dataset of one data instance; (b) the corresponding $f(x_1; w) - y_1$; (c) the corresponding $E(w) = (f(x_1; w) - y_1)^2$.

Figure 1 (a) shows a dataset of just one data instance: $(x_1, y_1) = (2, 2)$. Now we want to fit a nonlinear curve $f(x; w) = e^{wx}$ to the dataset, using the sum of square error (SSE): $E(w) = (f(x_1; w) - y_1)^2$. Figure 1 (b) and (c) show the curve of $f(x_1; w) - y_1$ w.r.t. w and the curve of $E(w)$ w.r.t. w, respectively.

1.1 Gauss-Newton method [9 points in total]

Let us use the Gauss-Newton method introduced in the lectures to find the solution $w^\star$ that minimizes $E(w)$. Note that, the optimal solution is 0.347. Since the Gauss-Newton method is an iterative method, the solution depends on the number of iteration.

1. Begin with the initialization $\hat{w}^{(1)} = 1.5$, perform three iterations of Gauss-Newton to get $\hat{w}^{(4)}$. What is $\hat{w}^{(4)}$ (a numerical value)? [3 point]

2. Begin with the initialization $\hat{w}^{(1)} = 0.0$, perform three iterations of Gauss-Newton to get $\hat{w}^{(4)}$. What is $\hat{w}^{(4)}$ (a numerical value)? [3 point]

3. Begin with the initialization $\hat{w}^{(1)} = -1.0$, perform three iterations of Gauss-Newton to get $\hat{w}^{(4)}$. What is $\hat{w}^{(4)}$ (a numerical value)? [3 point]

Your answers must contain numerical values. For example, 0.500 is allowed but 1/2 is not allowed. Please round your answer to the third decimal. For example, 1.333333 becomes 1.333; -1.333333 becomes -1.333.

Do you see that iterative methods are sensitive to the initialization? [0 point]

1.2 Gradient descent [9 points in total]

Let us now apply the gradient descent (GD) method introduced in the lectures to find the solution $w^\star$ that minimizes $E(w)$. Since GD is an iterative method, the solution depends on the number of iteration.

For our problem where w is one-dimensional, given the current guess $\hat{w}^{(t)}$, GD performs the following update:

$$\hat{w}^{(t+1)} \leftarrow \hat{w}^{(t)} - \eta \frac{\mathrm{d}E}{\mathrm{d}w}(\hat{w}^{(t)}), \tag{1}$$

where $\frac{\mathrm{d}E}{\mathrm{d}w}(\hat{w}^{(t)})$ is the derivative computed at $\hat{w}^{(t)}$ and η is the step size or learning rate.

1. Begin with the initialization $\hat{w}^{(1)} = 1.5$, perform three iterations of GD (with $\eta = 0.1$) to get $\hat{w}^{(4)}$. What is $\hat{w}^{(4)}$ (a numerical value)? [3 point]

2. Begin with the initialization $\hat{w}^{(1)} = 0.0$, perform three iterations of GD (with $\eta = 0.1$) to get $\hat{w}^{(4)}$. What is $\hat{w}^{(4)}$ (a numerical value)? [3 point]

3. Begin with the initialization $\hat{w}^{(1)} = 1.5$, perform three iterations of GD (with $\eta = 0.001$) to get $\hat{w}^{(4)}$. What is $\hat{w}^{(4)}$ (a numerical value)? [3 point]

Do you see that iterative methods are sensitive to the initialization and the step size? [0 point]

2 Regression, Gauss-Newton, and gradient descent: Part-2 [15 points]

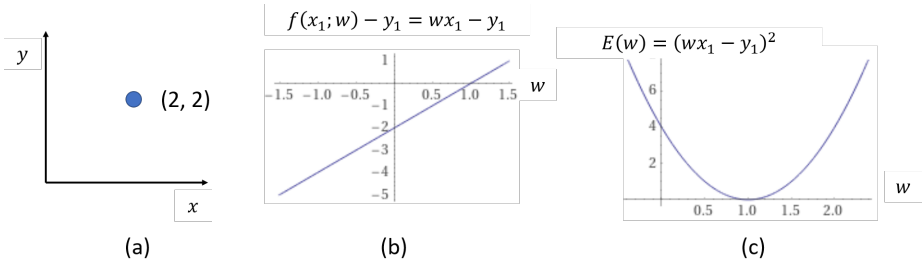


Figure 2: (a) A dataset of one data instance; (b) the corresponding $f(x_1; w) - y_1$; (c) the corresponding $E(w) = (f(x_1; w) - y_1)^2$.

Following Question 1, now let us consider a simpler problem: to fit a linear line $f(x; w) = wx$ to the same data, again using the sum of square error (SSE): $E(w) = (f(x_1; w) - y_1)^2$. Figure 2 (a) shows the same data as Figure 1, but Figure 2 (b) and (c) show the curve of $f(x_1; w) - y_1$ w.r.t. w and the curve of $E(w)$ w.r.t. w, respectively, using $f(x; w) = wx$.

2.1 Gauss-Newton method [6 points in total]

Let us again use the Gauss-Newton method to find the solution $w^\star$ that minimizes $E(w)$ in Figure 2. Note that, the optimal solution is 1.000.

1. Begin with the initialization $\hat{w}^{(1)} = 1.5$, perform just **one** iteration of Gauss-Newton to get $\hat{w}^{(2)}$. What is $\hat{w}^{(2)}$ (a numerical value)? [3 point]

2. Begin with the initialization $\hat{w}^{(1)} = 0.0$, perform just **one** iteration of Gauss-Newton to get $\hat{w}^{(2)}$. What is $\hat{w}^{(2)}$ (a numerical value)? [3 point]

Do you see any difference to the observations in Question 1.1? [0 point]

2.2 Gradient descent [9 points in total]

Let us now apply gradient descent (GD) to find the solution $w^\star$ that minimizes $E(w)$. For our problem where w is one-dimensional, given the current guess $\hat{w}^{(t)}$, GD performs the following update:

$$\hat{w}^{(t+1)} \leftarrow \hat{w}^{(t)} - \eta \frac{\mathrm{d}E}{\mathrm{d}w}(\hat{w}^{(t)}), \tag{2}$$

where $\frac{\mathrm{d}E}{\mathrm{d}w}(\hat{w}^{(t)})$ is the derivative computed at $\hat{w}^{(t)}$ and η is the step size or learning rate.

1. Begin with the initialization $\hat{w}^{(1)} = 1.5$, perform three iterations of GD (with $\eta = 0.1$) to get $\hat{w}^{(4)}$. What is $\hat{w}^{(4)}$ (a numerical value)? [3 point]
2. Begin with the initialization $\hat{w}^{(1)} = 0.0$, perform three iterations of GD (with $\eta = 0.1$) to get $\hat{w}^{(4)}$. What is $\hat{w}^{(4)}$ (a numerical value)? [3 point]
3. Begin with the initialization $\hat{w}^{(1)} = 1.5$, perform three iterations of GD (with $\eta = 1.0$) to get $\hat{w}^{(4)}$. What is $\hat{w}^{(4)}$ (a numerical value)? [3 point]

Can GD obtain the optimal solution in a few iterations? [0 point]

3 Naive Bayes and MLE [23 points]

	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$	$\boldsymbol{x}_7$	$\boldsymbol{x}_8$	$\boldsymbol{x}_9$	$\boldsymbol{x}_{10}$
$\boldsymbol{x}_i[1]$	1	1	1	1	0	0	0	0	0	1
$\boldsymbol{x}_i[2]$	1.0	2.0	2.5	3.5	4.0	5.0	0.5	1.0	2.0	2.5

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}
0	0	0	0	0	0	1	1	1	1

Figure 3: A two-dimensional labeled dataset with 10 data instances.

Figure 3 gives a labeled dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ of 10 data instances. Each $\boldsymbol{x}_i$ is two-dimensional: the first dimension $\boldsymbol{x}_i[1]$ is binary (i.e., $\{0, 1\}$); the second dimension $\boldsymbol{x}_i[2]$ is a real number. The label y_i is binary (i.e., either class 0 or class 1).

Denote by X the two-dimensional random variable of data instances and Y the binary random variable of class labels, you are to construct the Naive Bayes classifier to predict the class label $\hat{y}$ of a future data instance $X = \boldsymbol{x}$

$$\begin{aligned}\hat{y} &= \arg\max_{c\in\{0,1\}} p(Y = c|X = \boldsymbol{x}) \\ &= \arg\max_{c\in\{0,1\}} p(Y = c) \times \prod_{d=1}^{2} p(X[d] = x[d]|Y = c).\end{aligned}$$

You will begin by estimating the parameters of

- $p(Y)$
- $p(X[d]|Y = c)\ \forall c \in \{0, 1\}, d \in \{1, 2\}$

from the labeled dataset, using **Maximum Likelihood Estimation (MLE)**. Note that, $p(Y)$ is a Bernoulli distribution; $p(X[1]|Y = c)$ is a Bernoulli distribution of $X[1]$; $p(X[2]|Y = c)$ is a Gaussian distribution of $X[2]$.

3.1 Prior distributions [3 points]

What is $p(Y = 1)$? The answer is a real number.

3.2 Conditional distributions [6 points]

What is $p(X[1] = 1|Y = 0)$?

What is $p(X[1] = 1|Y = 1)$?

Each answer is a real number.

3.3 Conditional distributions [6 points]

What is $p(X[2] = 3.0|Y = 0)$?

What is $p(X[2] = 3.0|Y = 1)$?

Each answer is a real number.

3.4 Naive Bayes [8 points]

Given $\boldsymbol{x} = [0, 3.0]^\top$ (i.e., $x[1] = 0$ and $x[2] = 3.0$), what is the prediction $\hat{y}$?

Given $\boldsymbol{x} = [1, 0.5]^\top$ (i.e., $x[1] = 1$ and $x[2] = 0.5$), what is the prediction $\hat{y}$?

Each answer is either 0 or 1.

4 Probabilistic Graphical models (PGMs) [22 points]

4.1 Joint distribution [4 points]

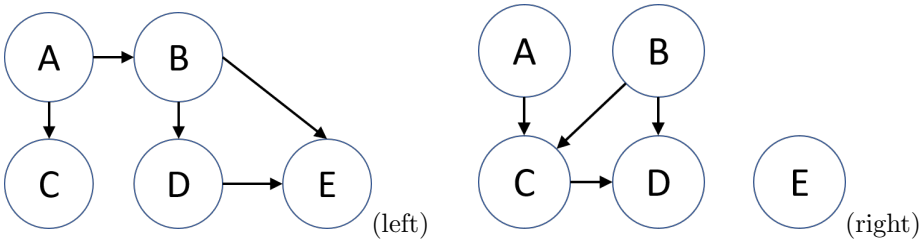


Figure 4: Two probabilistic graphical models (PGMs).

Figure 4 shows two PGMs of five random variables A, B, C, D, and E. A joint distribution such as $p(A, B, C, D, E)$ from a PGM can be decomposed into the form of:

$$\prod_{X \in \{A,B,C,D,E\}} p(X|\text{parents}(X))$$

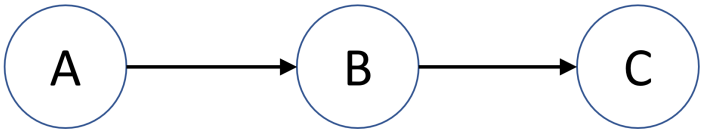
(a) Decompose $p(A, B, C, D, E)$ for the left PGM in the above-mentioned form. The expected answer is an expression that solely comprises of probabilities over one or more of the random variables A, B ,C, D, E.

(b) Decompose $p(A, B, C, D, E)$ for the right PGM in the above-mentioned form. The expected answer is an expression that solely comprises of probabilities over one or more of the random variables A, B ,C, D, E.

4.2 PGMs from joint distribution decomposition [2 points]

Please draw the PGM of five random variables A, B, C, D, and E, given that the joint distribution $p(A, B, C, D, E) = p(A)p(B)p(C)p(D|A, B)P(E|D, C)$.

4.3 Inference [16 points]



$A =$	$p(A)$
a	0.8
$\neg a$	0.2

$B =$	$A =$	$p(B\|A)$
b	a	0.6
$\neg b$	a	0.4
b	$\neg a$	0.3
$\neg b$	$\neg a$	0.7

$C =$	$B =$	$p(C\|B)$
c	b	0.5
$\neg c$	b	0.5
c	$\neg b$	0.1
$\neg c$	$\neg b$	0.9

Figure 5: A probabilistic graphical models (PGM) of three random variables and the corresponding probability terms.

Figure 5 shows a PGM of three binary random variables A, B, and C, together with the corresponding probability terms. Each binary random variable has two outcomes. Please compute the following probabilities (answer must be a floating point decimal) and answer the following questions with a response of Yes or No.

1. $p(B = b)$ [2 points]
2. $p(C = \neg c)$ [2 points]
3. $p(A = a, C = \neg c)$ [2 points]
4. Is $p(A = a, C = \neg c)$ equal to $p(A = a)p(C = \neg c)$? [2 point]
5. $p(A = a|B = b)$ [2 points]
6. $p(A = a, C = \neg c|B = b)$ [2 points]
7. Is $p(A = a, C = \neg c|B = b)$ equal to $p(A = a|B = b)p(C = \neg c|B = b)$? [2 point]
8. $p(A = a|C = \neg c)$ [2 points]

Hint:

- $P(C = \neg c) = P(C = \neg c, B = b) + P(C = \neg c, B = \neg b) = P(C = \neg c|B = b)p(B = b) + P(C = \neg c|B = \neg b)p(B = \neg b)$
- $P(A = a, C = \neg c) = P(A = a, C = \neg c, B = b) + P(A = a, C = \neg c, B = \neg b)$

5 Conditional Independence [10 points]

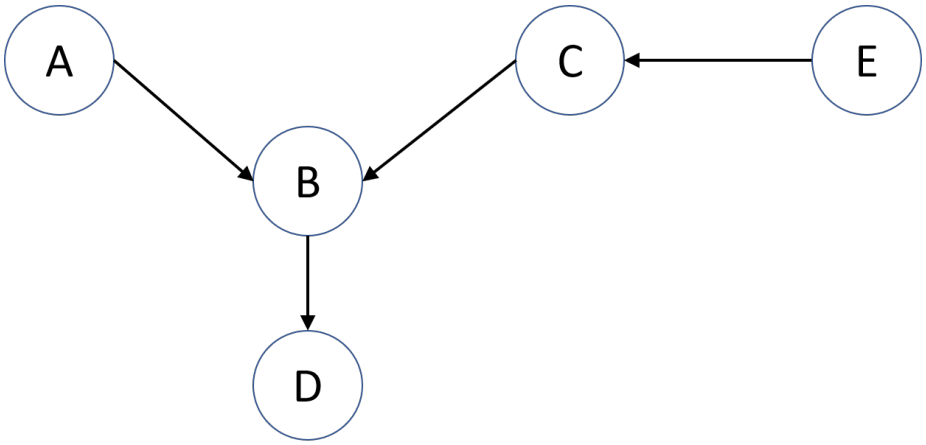


Figure 6: A probabilistic graphical models (PGM) of five random variables.

Figure 6 shows a PGM of five random variables A, B, C, D, and E. Please answer the following questions.

1. Is $A \perp E$ (i.e., $p(A, E) = p(A)P(E)$)?
2. Is $A \perp D|B$ (i.e., $p(D|A, B) = p(D|B)$)?
3. Is $A \perp E|B$ (i.e., $p(A, E|B) = p(A|B)p(E|B)$)?
4. Is $A \perp E|B, C$ (i.e., $p(A, E|B, C) = p(A|B, C)p(E|B, C)$)?
5. Is $A \perp C|D$ (i.e., $p(A, C|D) = p(A|D)p(C|D)$)?

6 K-means [12 points]

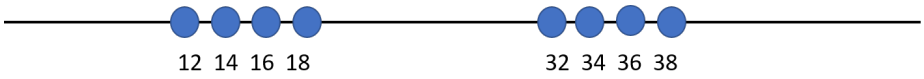


Figure 7: A one-dimensional dataset of 8 data instances

Figure 7 shows a dataset of 8 data instances, each of them is one-dimensional.

1. Please perform K-means for 10 iterations, given $K = 2$ and $c_1^{(t)} = 25$ and $c_2^{(t)} = 41$ when $t = 1$ (i.e., initialization). $c_1^{(t)}$ and $c_2^{(t)}$ represent the $K = 2$ centers in the beginning of the t-th iteration. Please use Euclidean distance. What will be $c_1^{(11)}$ and $c_1^{(11)}$? [6 points]

2. Please perform K-means for 10 iterations, given $K = 2$ and $c_1^{(t)} = 25$ and $c_2^{(t)} = 50$ when $t = 1$ (i.e., initialization). $c_1^{(t)}$ and $c_2^{(t)}$ represent the $K = 2$ centers in the beginning of the t-th iteration. Please use Euclidean distance. What will be $c_1^{(11)}$ and $c_1^{(11)}$? [6 points]