Computer Science

CSC380: Principles of Data Science

Final review

Chicheng Zhang

Announcements 5/7

- We will continue to hold office hours until final!
  - Office hour next Tuesday 5/13 moved to 10am
- Project submission
  - We created a ‘project code’ entry in gradescope
  - Submit your data there as well

Logistics

- Time: May 13 3:30-5:30pm
- Place: WSEL200W (here)

- You are welcome to:
    - Bring calculators
    - Bring a letter-size “cheatsheet” (formulas, examples, pictures)

- Scope:
    - topics covered after midterm

Logistics

- Problem types:
  - True/False questions (8 questions, 16pts)
    - We also ask you to provide brief justifications
  - Multiple-choice questions (6 questions, 24pts)
    - There can be multiple correct choices
    - Select all correct choices to get full credit
  - Free-form questions (about 9 questions, total 60 pts)
    - We don't expect them to be calculation heavy (mainly assessing understanding of basic concepts / methods)
    - We expect answers with justifications (steps)

General tips

- Review:
  - Lecture slides
  - Homeworks
  - Quizzes
  - Practice problem set
- We will go over some example questions below
  - We will also release answer keys soon
- Let us know if anything is unclear
  - We are here to help (in office hours or online)

Q1

1. Suppose that expectation of a random variable X is $E(X) = 5$. What is $E(3X - 5)$?

- Linearity of expectation

- $\mathrm{E}[3X - 5] = \mathrm{E}[3X] - 5 = 3\,\mathrm{E}[X] - 5 = 10$

Q4

4. Suppose that a random X can take each of the five values $-2, 0, 1, 3, 4$ with equal probability. Compute the variance of X using the formula $Var(X) = E(X^2) - (E(X))^2$. Also compute the variance of $Y = 2X - 10$.

... after some work, we found that $\text{Var}(X) = 4.56$

How can we find $\text{Var}(Y) = \text{Var}(2X - 10)$?

$$
\begin{aligned}
\text{Var}(2X - 10) &= \text{Var}(2X) \\
&= 2^2\,\text{Var}(X) = 18.24
\end{aligned}
$$

Variance is preserved with constant shifts

Var(aX) = a^2 Var(X)

Q3

3. Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Let $\hat{\mu} := \frac{1}{n}\sum_{i=1}^{n} X_i$. What distribution does $\hat{\mu}$ follow? Fully specify the parameters of the distribution. Show your work and reasoning.

- Actually, this problem needs some extra conditions to be solvable..
- We need to assume that $X_1, \ldots, X_n$ are independent
- What is the type of distribution of $\hat{\mu}$?
  - Gaussian
- What specific Gaussian distribution does $\hat{\mu}$ follow?
  - Figure out its mean and variance
  - See quiz 7

Q5

5. Suppose $X \sim \text{Binomial}(10, 0.6)$, and $Y \sim \text{Binomial}(8, 0.6)$, and X and Y are independent. Find the distribution of $Z = X + Y$. What is its variance? (Hint: it may be useful to think of X as a sum of independent Bernoulli random variables.)

- This is a tricky question.. (such questions are rare in exam)
- Idea 1:
  - Figure out all values Z can take (0, 1, .., 18)
  - Find $\mathrm{P}(Z = 0) = P(X = 0, Y = 0)$,

    $\mathrm{P}(Z = 1) = P(X = 0, Y = 1) + P(X = 1, Y = 0)$

    ...

  - Tons of calculations

Q5

5. Suppose $X \sim \text{Binomial}(10, 0.6)$, and $Y \sim \text{Binomial}(8, 0.6)$, and X and Y are independent. Find the distribution of $Z = X + Y$. What is its variance? (Hint: it may be useful to think of X as a sum of independent Bernoulli random variables.)

- This is a tricky question.. (such questions are rare in exam)
- Idea 2:
  - Think about these random variables *physically*
  - X = #shots made in the first 10 trials (with success prob. 0.6)
  - Y = #shots made in the next 8 trials (with success prob. 0.6)

  - X+Y = #shots made in a total of 18 trials (with success prob. 0.6)
  - Thus, $\sim \text{Binomial}(18, 0.6)$

Q9

9. Suppose we have placed two advertisements next to each other in a website. A user can either click both, click one of them, or not click at all. Let $A \in \{1,0\}$ and $B \in \{1,0\}$ be the random variables indicating whether each ad is clicked (1) or not(0). They follow the following joint probability table.

	$B=1$	$B=0$
$A=1$	$1/8$	$3/8$
$A=0$	$3/8$	$1/8$

(i) Which distribution does the random variable A follow? Specify the parameter of the distribution as well.

(ii) Are A and B independent? Justify your answer.

1. From joint distribution to marginal distribution: marginalization
2. How to tell if A and B are independent?

   4 equalities need to hold: $P(A=0, B=0) = P(A=0)P(B=0)$ ….

Q19

19. Associate each plot below with one of the correlation coefficient values: $1, 0.8, 0.4, 0, -0.4, -0.8, -1$.

Figure 1: 7 samples of (X, Y) with different joing distributions.

- Correlation coefficient ρ is always in [-1, 1]
- $\rho =$ -1 / +1 : X, Y are perfectly negatively / positively correlated

Example true/false question

- True or false: expectation of product of two random variables X, Y is equal to the product of expectations of X, Y
- Is $\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]$?
- Not necessarily – the above is equivalent to $\mathrm{Cov}(X, Y) = 0$

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

- this is false when X, Y are positively / negatively correlated

Q12

12. Suppose that we observe 4 data points $S = (x_1, x_2, x_3, x_4) = (3, 0, 1, 2)$ from Binomial$(3, \theta)$ and wish to estimate θ.
(a) Compute the log-likelihood function $\ln L(\theta)$.
(b) Compute the value of the maximum likelihood estimator for θ. You can use the fact that $\ln L(\theta)$'s only stationary point is its maximizer.

- Again this is a harder question (maybe better for HWs)
- Step 1: write down the log-likelihood

$$\ln L(\theta) = \sum_{i=1}^{4} \ln f(x_i; \theta)$$

$f(x_i; \theta)$: PMF of Binomial(3,θ) on example x_i

$$= \binom{3}{x_i} \theta^{x_i} (1-\theta)^{3-x_i}$$

Taking the natural log, it is $\ln \binom{3}{x_i} + x_i \ln \theta + (3 - x_i) \ln(1-\theta)$

Q12

- Step 2: simplify the log-likelihood

$$\ln L(\theta) = \sum_{i=1}^{4} x_i \ln\theta + (3 - x_i)\ln(1-\theta) + \ln\binom{3}{x_i}$$

$$= 6\ln\theta + 6\ln(1-\theta) + \text{ constant}$$

Step 3: find the parameter that maximize the log-likelihood

as given by the hint, finding stationary point of $\ln L(\theta)$ suffices

An example multiple-choice question

Q2: In least-squares linear regression, which of the following are true about the trained model?

- ○ It fits a linear function to input-output pairs
- ○ It always achieves zero training loss
- ○ It minimizes the average square loss on the training set
- ○ It assumes output labels are binary

- Answer: 1, 3
- Points will be calculated based on the correctness of each check box

Q14

14. A random sample of n items is to be taken from a distribution with mean μ and standard deviation σ. Use the central limit theorem to determine the smallest number of items n that must be taken in order to satisfy

$$P(|\overline{X}_n - \mu| \leq \frac{\sigma}{4}) \geq 0.99$$

approximately. You may make use of the following outputs from scipy.stats: $norm.ppf(0.995) = 2.58$, $norm.ppf(0.95) = 1.64$, $norm.cdf(0.995) = 0.84$, $norm.cdf(0.95) = 0.83$.

- General methodology: represent the probability using some RV whose distribution we know about (e.g. N(0,1))
- Here, we can consider the Z-statistic:

$$Z = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \sim N(0,1)$$

Holds approximately, by CLT

Q14

$$P(|\overline{X}_n - \mu| \leq \frac{\sigma}{4}) \geq 0.99$$

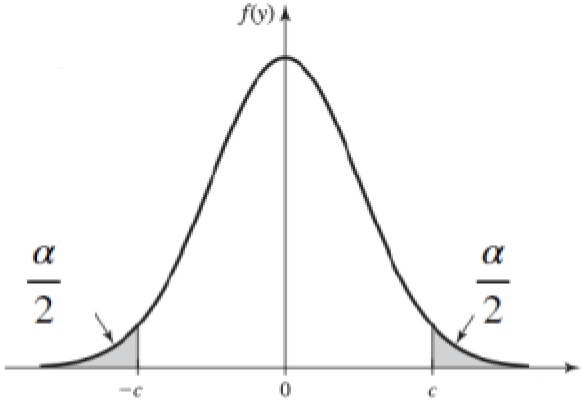
- Can we write it in terms of Z?

$$Z = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$$

$$P\left(|Z| \leq \frac{\sqrt{n}}{4}\right) \geq 0.99$$

- Since norm.ppf(0.995) = 2.58

$$P(|Z| \leq 2.58) \geq 0.99$$



- We should pick n such that $\frac{\sqrt{n}}{4} \geq 2.58 \Rightarrow n \geq 106.5$

Q24

24. Compute the accuracy and F-score values for the following scenario: 10% of the items are positive and the rest are negative. Suppose we are using a random classifier that classifies the items as positive with 0.6 probability.

$f(x)$: classifier's prediction $\qquad$ y: example's true label

Random classifier: $f(x)$ and y and independent

Accuracy: $\mathrm{P}(f(x) = y) = P(f(x) = 1, y = 1) + P(f(x) = 0, y = 0)$

independence, = 0.6 * 0.1 $\qquad$ independence, = 0.4 * 0.9

= 0.06 + 0.36 = 0.42

Q24

24. Compute the accuracy and F-score values for the following scenario: 10% of the items are positive and the rest are negative. Suppose we are using a random classifier that classifies the items as positive with 0.6 probability.

$f(x)$: classifier's prediction $\quad$ y: example's true label

Random classifier: $f(x)$ and y and independent

F-score: $\frac{2}{Recall^{-1}+Precision^{-1}}$

Recall: $P(\, f(x) = 1 \mid y = 1\,) = P(f(x) = 1) = 0.6$

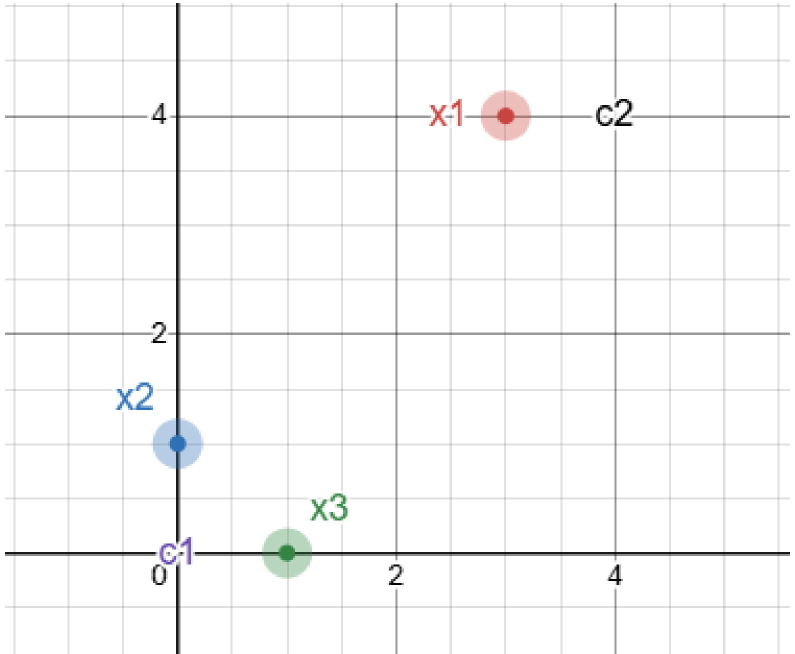
Precision: $P(\, y = 1 \mid f(x) = 1\,) = P(y = 1) = 0.1$

F-score = 0.171

Q29

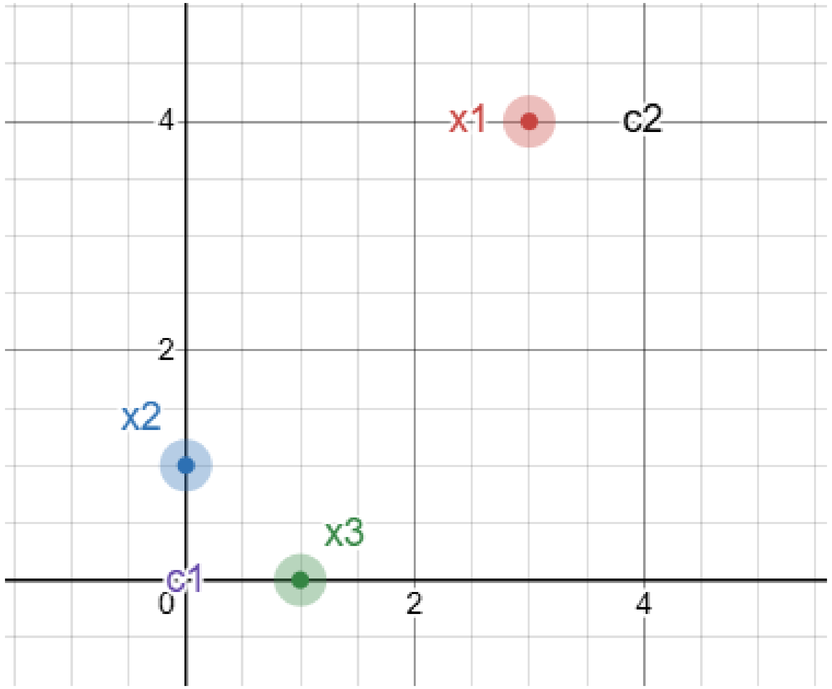
29. Suppose that we have the following 3 data points $x_1 = (3, 4), x_2 = (0, 1), x_3 = (1, 0)$ Starting from the initial centroids $c_1 = (0, 0)$ and $c_2 = (4, 4)$, run the k-means clustering algorithm until the centroids don't move anymore. State your final clustering result (i.e., state which points are in the same cluster)

- First, we always recommend drawing a 2D picture for such geometric problems



Q29

29. Suppose that we have the following 3 data points $x_1 = (3, 4), x_2 = (0, 1), x_3 = (1, 0)$ Starting from the initial centroids $c_1 = (0, 0)$ and $c_2 = (4, 4)$, run the k-means clustering algorithm until the centroids don't move anymore. State your final clustering result (i.e., state which points are in the same cluster)



k-means repeat:

step 1. re-assigning the clusters

step 2. update the centroids

Iteration 1:

step 1: x_2, x_3 goes to cluster 1, x_1 goes to cluster 2

why?

Q29

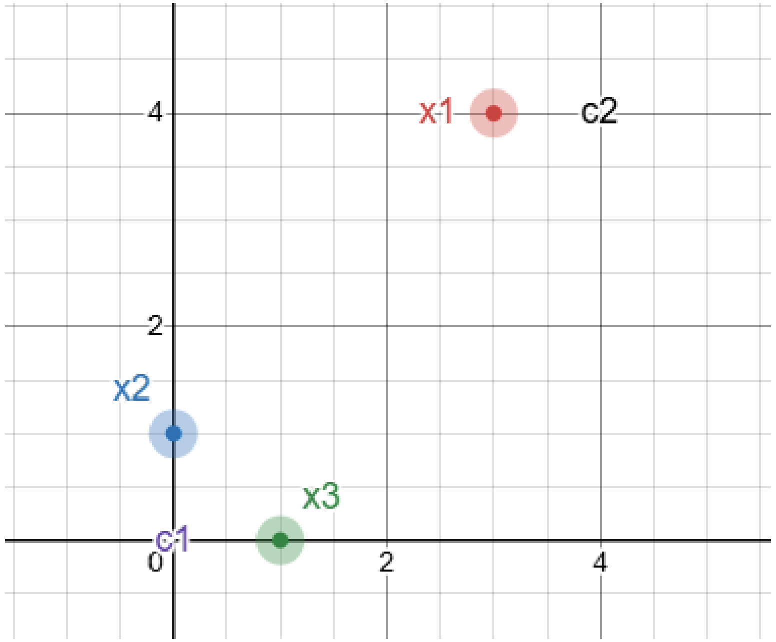
- Iteration 1:

  step 1: x_2, x_3 goes to cluster 1, x_1 goes to cluster 2

  x_2 is closer to c_1 than c_2

  How to verify this rigorously?

  $$||x_2 - c_1|| = ||\,(0,1) - (0,0)|| = ||(0,1)||$$
  $$= \sqrt{0^2 + 1^2} = 1$$
  $$||x_2 - c_2|| = ||\,(0,1) - (4,4)|| = 5$$



Also useful in nearest neighbor classification

  step 2: new centroids:

  c_1 updated to $(0.5,0.5)$, c_2 updated to (3,4)

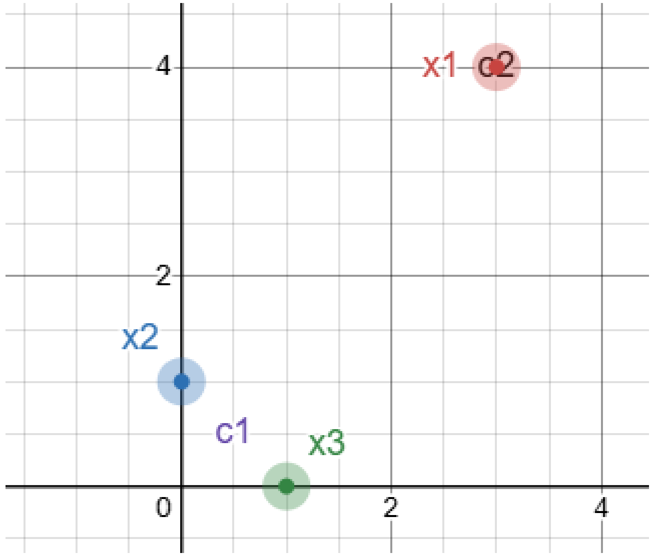
Q29

29. Suppose that we have the following 3 data points $x_1 = (3,4), x_2 = (0,1), x_3 = (1,0)$ Starting from the initial centroids $c_1 = (0,0)$ and $c_2 = (4,4)$, run the k-means clustering algorithm until the centroids don't move anymore. State your final clustering result (i.e., state which points are in the same cluster)

Iteration 2:

step 1: x_2, x_3 goes to cluster 1, x_1 goes to cluster 2

step 2: new centroids: c_1 updated to (0.5,0.5), c_2 updated to (3,4)



The centroids don't move any more.
We are done!

Hope you all excel in finals!