IFT 3395/6390 (6390: GRAD) Theoretical Homework 1

Akshay Singh Rana

TOTAL POINTS

60 / 74

QUESTION 1

Probability warm-up: conditional probabilities and Bayesrule 5 pts

1.1 a 1 / 1

✓ **- 0 pts Correct**

- 1 pts This do not define P(X|Y)

- 0 pts Copy pasted an entire paragraph with the answer somewhere inside

- 0.5 pts switched X and Y

- 1 pts no answer or no page selected

1.2 b 0 / 1

- 0 pts Correct

✓ **- 1 pts wrong reasoning/justification**

- 0.5 pts calculation error

- 1 pts has not answered or selected page

- 0.5 pts justification lacks key elements

- 0.25 pts notation error

- 0.25 pts small error in the justification or lack of clarity

1.3 c 1 / 1

✓ **- 0 pts Correct**

- 0.5 pts 1 wrong expression

- 1 pts wrong expression

- 1 pts question not understood

- 1 pts no answer or no page selected

1.4 d 0 / 1

- 0 pts Correct

✓ **- 1 pts incorrect**

- 0.5 pts confusing proof

- 1 pts no answer or no page selected

1.5 e1 0 / 0

- 0 pts Correct

✓ **- 0 pts does not count in grade**

1.6 e2 1 / 1

✓ **- 0 pts Correct**

- 1 pts did not answer the question

- 1 pts wrong result

- 0.5 pts calculation error

- 0.25 pts did not make the calculation

- 0.25 pts rounding error

- 0 pts no answer or no page selected

QUESTION 2

Bag of words and single topic model 12 pts

2.1 a 2 / 2

✓ **- 0 pts Correct**

- 2 pts Wrong

- 2 pts No selected pages

2.2 b 2 / 2

✓ **- 0 pts Correct**

- 2 pts Wrong

- 2 pts No pages selected

- 2 pts Answer not on selected page

2.3 c 2 / 2

✓ **- 0 pts Correct**

- 0.5 pts Reduce to 9/1000

- 2 pts Wrong

- 2 pts No pages selected

2.4 d 2 / 2

✓ **- 0 pts Correct**

- **2 pts** Wrong
- **1 pts** Reduce to 10/13
- **2 pts** Answer not on selected page
- **0.5 pts** Bad rounding (should be 0.77)
- **2 pts** Write final answer not equation

2.5 e 2 / 2

✓ - **0 pts** **Correct**
- **2 pts** Wrong
- **1 pts** Reduce to 121/13000
- **2 pts** Write answer not equation
- **2 pts** Answer not on selected page
- **1 pts** Bad rounding

2.6 f 2 / 2

✓ - **0 pts** **Correct**
- **1 pts** No document topic probabilities
- **1 pts** Wrong conditional probabilities
- **2 pts** Wrong
- **2 pts** Answer not on selected page
- **0.5 pts** No need for sampling
- **0.5 pts** Be more specific

QUESTION 3

Maximum likelihood estimation 5 pts

3.1 a 1 / 1

✓ - **0 pts** **Correct**
- **0.5 pts** Definition of distribution not rigorous enough
- **1 pts** Did not express in terms of f(xi)'s
- **1 pts** no answer or no page selected

3.2 b 3.5 / 4

✓ - **0 pts** **Correct**
✓ - **0.5 pts** **Definition of the likelihood/or loglikelihood function not rigorous enough**
- **1.5 pts** Did not define/or derive properly the likelihood/ or loglikelihood function
- **1.5 pts** Insufficient proof: did not explain/ or incorrectly explained the behavior of the likelihood or the loglikelihood in terms of theta
- **1 pts** Insufficient proof: did not explain/ or incorrectly explained the domain of theta in likelihood
- **0.5 pts** The xi's are not necessarily ordered
- **4 pts** no answer or no page selected

QUESTION 4

4 Maximum likelihood estimation 2 7 / 10

- **0 pts** Correct
- **2 pts** Calculation mistake
- **5 pts** Did not derive likelihood properly or did not derive likelihood for n samples

✓ - **2 pts** **Did not prove stationary point was a maxima**

✓ - **1 pts** **Did not mention log(x) is a monotonic/increasing function to prove the use of log**
- **1 pts** Did not simplify expression enough
- **10 pts** no answer or no page selected

QUESTION 5

k-nearest neighbors 10 pts

5.1 a 1 / 4

- **0 pts** Correct
- **4 pts** no answer
- **1.5 pts** reasoning error / answer not detailed enough

✓ - **3 pts** **only traces of right answer**
- **4 pts** wrong
- **0.5 pts** you haven't selected all pages corresponding to your answer
- **0 pts** ALMOST ILLEGIBLE !
- **0 pts** DUP 1?
- **2 pts** k can't be 2n+1
- **0 pts** DUP 2?
- **4 pts** no pages selected

5.2 b 0 / 2

- **0 pts** Correct

✓ - **2 pts** **wrong**
- **1 pts** Starting from the desired result and going through implications to something that's true is not

how proofs work in math. If you want to do this, you need to replace your implications by a sentence like "it suffices to prove that..." or use equivalences if possible. If you haven't written anything between equations in different lines, I assume it's an implication !

- **1.5 pts** you didn't generalize to all k
- **2 pts** wrong induction
- **2 pts** no answer
- **0.5 pts** "at least n" is wrong
- **0.5 pts** we don't *have to* prove Pn(e) is an increasing fct wrt k
- **1 pts** k can't be 2
- **0.5 pts** you didn't conclude...
- **2 pts** no pages selected

5.3 C 3 / 4

- **0 pts** Correct
- **4 pts** wrong
- **0.5 pts** what about k=1, 3, 5 ?
- **0.5 pts** a sqrt(n) is not necessarily an integer and/or odd

✓ **- 1 pts error, (check the box)**

- **0.5 pts** you can't talk about lim P_n(e) or lim 1/2^n sum(...) before proving their existence
- **0.5 pts** a limit can't be equal to something that depends on n
- **4 pts** no pages selected
- **4 pts** no answer
- **3 pts** traces of answer only
- **0.5 pts** correct but needs more rigor
- **2 pts** errors (check the boxes)
- **1 pts** error (check the comment)
- **0.5 pts** a few mistakes in the induction, but mostly ok
- **1.5 pts** error, check the comment
- **0 pts** DUP 1?
- **0 pts** DUP 2?
- **0.5 pts** n/2 not necessarily integer
- **0.5 pts** error (check comment)
- **0.5 pts** limits are not "equivalent to each other"
- **2 pts** you just showed that P_n(e) < 1, while your proof strategy can be quickly adapted to show that P_n(e) goes to 0
- **1 pts** it doesn't matter that the denominator is O(2^n), what matters is that it's Omega(2^n)
- **1 pts** You just proved that a the cdf of a certain binomial RV evaluated at a certain point converges to 0. You didn't make the link with P_n(e)
- **0.5 pts** you didn't select all the pages corresponding to your answer

1 lacks a sum here

QUESTION 6

6 Gaussian Mixture 10 / 10

✓ **- 0 pts Correct**

- **1 pts** result is not simplified
- **2 pts** calculation error : sigma1 and sigma2 exchanged
- **2 pts** calculation error : additional or missing term
- **10 pts** wrong reasoning
- **2 pts** calculation error
- **2 pts** notation error / lack of rigor
- **7.5 pts** serious mathematical error
- **2 pts** calculation error : wrong sign
- **7.5 pts** serious calculation error
- **2 pts** result is not simplified at all
- **5 pts** unnecessary reasoning error / confusion
- **10 pts** no answer or no page selected
- **1 pts** illegible
- **10 pts** no justification
- **5 pts** unsufficient justification

QUESTION 7

Practical Report 22 pts

7.1 practical report Q5 7 / 7

✓ **- 0 pts Correct**

- **2.5 pts** error in one of the curves
- **5 pts** error in both curves
- **2 pts** no discussion / comment
- **1 pts** incomplete discussion / comment
- **1 pts** one of the curves is not totally right

- **7 pts** no pages selected
- **7 pts** no answer
- **0.5 pts** it is literally mentioned in the homework : "in the same plot"
- **0 pts** Click here to replace this description.
- **2 pts** wrong axes

7.2 practical report Q7 5 / 5

✓ - **0 pts** **Correct**
- **0 pts** no variation in h
- **0 pts** soft parzen < hard parzen
- **0 pts** weird time curves
- **2.5 pts** wrong interpretation (soft)
- **0 pts** ignored practical running time
- **5 pts** false reasoning
- **5 pts** absurd results
- **2.5 pts** wrong interpretation (hard)
- **5 pts** page not selected or no answer
- **2 pts** hard parzen complexity is not proportional to neighbors
- **2.5 pts** does not explain dependency
- **3 pts** interpretation too superficial or does not match obs.
- **5 pts** does not answer the question
- **3 pts** partial error in interpretation
- **1.5 pts** error in theo. complexity calculation
- **2 pts** unclear justification
- **1 pts** hardly legible

7.3 practical report Q9 7.5 / 10

- **0 pts** Correct

✓ - **2.5 pts** **missing/wrong error bars**
- **4 pts** one of the curves is wrong
- **8 pts** both curves are wrong
- **2 pts** no comparison / discussion
- **2 pts** one of the curves is not exactly right
- **4 pts** both curves are not exactly right
- **10 pts** no pages selected
- **10 pts** no answer
- **2 pts** your plot should be a line plot
- **0 pts** DUP 1
- **0 pts** DUP 2

IFT-6390 Fundamentals of Machine Learning
Professor: Ioannis Mitliagkas

Homework 1 - Theoretical part

AKSHAY SINGH RANA

1. **Probability warm-up: conditional probabilities and Bayes rule** [5 points]

   (a) Give the definition of the conditional probability of a discrete random variable X given a discrete random variable Y.
   Answer. The conditional probability of a X is the probability of X given that we know the certain value of a discrete random variable Y.

   $$P(X = x|Y = y) = \frac{P(\{Y = y\} \cap \{X = x\})}{P(Y = y)}$$

   (b) Consider a biased coin with probability 2/3 of landing on heads and 1/3 on tails. This coin is tossed three times. What is the probability that exactly two heads occur (out of the three tosses) given that the first outcome was a head?
   Answer. 3 coin tosses will have 8 cases, but since the first toss is head and we need exactly two heads in three tosses, this leaves us with just two cases i.e. [HHT, HTH]

   $$\begin{aligned} Probability &= P(HHT) + P(HTH) \\ &= \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} + \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} \\ &= \frac{4}{27} + \frac{4}{27} \\ &= \frac{8}{27} \end{aligned}$$

   (c) Give two equivalent expressions of $P(X, Y)$:
      (i) as a function of $\mathbb{P}(X)$ and $\mathbb{P}(Y|X)$
      Answer.

      $$\mathbb{P}(X, Y) = \mathbb{P}(X|Y)\mathbb{P}(Y) \tag{1}$$

      (ii) as a function of $\mathbb{P}(Y)$ and $\mathbb{P}(X|Y)$
      Answer.

      $$\mathbb{P}(X, Y) = \mathbb{P}(Y|X)\mathbb{P}(X) \tag{2}$$

1.1 a **1 / 1**

✓ **- 0 pts Correct**

- 1 pts This do not define P(X|Y)

- 0 pts Copy pasted an entire paragraph with the answer somewhere inside

- 0.5 pts switched X and Y

- 1 pts no answer or no page selected

IFT-6390 Fundamentals of Machine Learning
Professor: Ioannis Mitliagkas

Homework 1 - Theoretical part

AKSHAY SINGH RANA

1. **Probability warm-up: conditional probabilities and Bayes rule** [5 points]

   (a) Give the definition of the conditional probability of a discrete random variable X given a discrete random variable Y.
   Answer. The conditional probability of a X is the probability of X given that we know the certain value of a discrete random variable Y.

   $$P(X = x|Y = y) = \frac{P(\{Y = y\} \cap \{X = x\})}{P(Y = y)}$$

   (b) Consider a biased coin with probability 2/3 of landing on heads and 1/3 on tails. This coin is tossed three times. What is the probability that exactly two heads occur (out of the three tosses) given that the first outcome was a head?
   Answer. 3 coin tosses will have 8 cases, but since the first toss is head and we need exactly two heads in three tosses, this leaves us with just two cases i.e. [HHT, HTH]

   $$\begin{aligned} Probability &= P(HHT) + P(HTH) \\ &= \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} + \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} \\ &= \frac{4}{27} + \frac{4}{27} \\ &= \frac{8}{27} \end{aligned}$$

   (c) Give two equivalent expressions of $P(X, Y)$:
      (i) as a function of $\mathbb{P}(X)$ and $\mathbb{P}(Y|X)$
      Answer.

      $$\mathbb{P}(X, Y) = \mathbb{P}(X|Y)\mathbb{P}(Y) \tag{1}$$

      (ii) as a function of $\mathbb{P}(Y)$ and $\mathbb{P}(X|Y)$
      Answer.

      $$\mathbb{P}(X, Y) = \mathbb{P}(Y|X)\mathbb{P}(X) \tag{2}$$

1.2 b **0 / 1**

- **0 pts** Correct

✓ - **1 pts** **wrong reasoning/justification**

- **0.5 pts** calculation error
- **1 pts** has not answered or selected page
- **0.5 pts** justification lacks key elements
- **0.25 pts** notation error
- **0.25 pts** small error in the justification or lack of clarity

IFT-6390 Fundamentals of Machine Learning
Professor: Ioannis Mitliagkas

Homework 1 - Theoretical part

AKSHAY SINGH RANA

1. **Probability warm-up: conditional probabilities and Bayes rule** [5 points]

   (a) Give the definition of the conditional probability of a discrete random variable X given a discrete random variable Y.
   Answer. The conditional probability of a X is the probability of X given that we know the certain value of a discrete random variable Y.

   $$P(X = x|Y = y) = \frac{P(\{Y = y\} \cap \{X = x\})}{P(Y = y)}$$

   (b) Consider a biased coin with probability 2/3 of landing on heads and 1/3 on tails. This coin is tossed three times. What is the probability that exactly two heads occur (out of the three tosses) given that the first outcome was a head?
   Answer. 3 coin tosses will have 8 cases, but since the first toss is head and we need exactly two heads in three tosses, this leaves us with just two cases i.e. [HHT, HTH]

   $$\begin{aligned} Probability &= P(HHT) + P(HTH) \\ &= \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} + \frac{2}{3} \times \frac{2}{3} \times \frac{1}{3} \\ &= \frac{4}{27} + \frac{4}{27} \\ &= \frac{8}{27} \end{aligned}$$

   (c) Give two equivalent expressions of $P(X, Y)$:
      (i) as a function of $\mathbb{P}(X)$ and $\mathbb{P}(Y|X)$
      Answer.

      $$\mathbb{P}(X, Y) = \mathbb{P}(X|Y)\mathbb{P}(Y) \tag{1}$$

      (ii) as a function of $\mathbb{P}(Y)$ and $\mathbb{P}(X|Y)$
      Answer.

      $$\mathbb{P}(X, Y) = \mathbb{P}(Y|X)\mathbb{P}(X) \tag{2}$$

1.3 C **1 / 1**

✓ **- 0 pts** **Correct**

- 0.5 pts 1 wrong expression

- 1 pts wrong expression

- 1 pts question not understood

- 1 pts no answer or no page selected

(d) Prove Bayes theorem:

$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(Y|X)\mathbb{P}(X)}{\mathbb{P}(Y)}.$$

Answer. Using 1 and 2 in the Bayes equation above.

$$\begin{aligned}\mathbb{P}(X|Y) &= \frac{\mathbb{P}(Y|X)\mathbb{P}(X|Y)\mathbb{P}(X,Y)}{\mathbb{P}(Y|X)\mathbb{P}(X,Y)} \\ &= \mathbb{P}(X|Y)\end{aligned}$$

(e) A survey of certain Montreal students is done, where 55% of the surveyed students are affiliated with UdeM while the others are affiliated with McGill. A student is drawn randomly from this surveyed group.

i. What is the probability that the student is affiliated with McGill?
Answer.

$$\begin{aligned}P(StudentfromMcGill) &= 1 - P(StudentfromUdeM) \\ &= 1 - 0.55 \\ &= 0.45\end{aligned}$$

ii. Now let's say that this student is bilingual, and you know that 80% of UdeM students are bilingual while 50% of McGill students are. Given this information, what is the probability that this student is affiliated with McGill ?

Answer. Using Bayes theorem

$$\begin{aligned}\mathbb{P}(Mcgill|Bilingual) &= \frac{\mathbb{P}(Bilingual|Mcgill)\mathbb{P}(Mcgill)}{\mathbb{P}(Bilingual)} \\ &= \frac{0.50 \times 0.45}{0.50 \times 0.45 \ + \ 0.8 \times 0.55} \\ &= \frac{0.225}{0.665} \\ &= 0.338\end{aligned}$$

2. **Bag of words and single topic model** [10 points] We consider a classification problem where we want to predict the topic of a document from a given corpus (collection of documents). The topic of each

1.4 d **0 / 1**

- **0 pts** Correct

✓ **- 1 pts incorrect**

- **0.5 pts** confusing proof
- **1 pts** no answer or no page selected

(d) Prove Bayes theorem:

$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(Y|X)\mathbb{P}(X)}{\mathbb{P}(Y)}.$$

Answer. Using 1 and 2 in the Bayes equation above.

$$\begin{aligned}\mathbb{P}(X|Y) &= \frac{\mathbb{P}(Y|X)\mathbb{P}(X|Y)\mathbb{P}(X,Y)}{\mathbb{P}(Y|X)\mathbb{P}(X,Y)} \\ &= \mathbb{P}(X|Y)\end{aligned}$$

(e) A survey of certain Montreal students is done, where 55% of the surveyed students are affiliated with UdeM while the others are affiliated with McGill. A student is drawn randomly from this surveyed group.

i. What is the probability that the student is affiliated with McGill?
Answer.

$$\begin{aligned}P(StudentfromMcGill) &= 1 - P(StudentfromUdeM) \\ &= 1 - 0.55 \\ &= 0.45\end{aligned}$$

ii. Now let's say that this student is bilingual, and you know that 80% of UdeM students are bilingual while 50% of McGill students are. Given this information, what is the probability that this student is affiliated with McGill ?

Answer. Using Bayes theorem

$$\begin{aligned}\mathbb{P}(Mcgill|Bilingual) &= \frac{\mathbb{P}(Bilingual|Mcgill)\mathbb{P}(Mcgill)}{\mathbb{P}(Bilingual)} \\ &= \frac{0.50 \times 0.45}{0.50 \times 0.45 \ + \ 0.8 \times 0.55} \\ &= \frac{0.225}{0.665} \\ &= 0.338\end{aligned}$$

2. **Bag of words and single topic model** [10 points] We consider a classification problem where we want to predict the topic of a document from a given corpus (collection of documents). The topic of each

1.5 e1 **0 / 0**

- **0 pts** Correct

✓ - **0 pts** **does not count in grade**

(d) Prove Bayes theorem:

$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(Y|X)\mathbb{P}(X)}{\mathbb{P}(Y)}.$$

Answer. Using 1 and 2 in the Bayes equation above.

$$\begin{aligned}\mathbb{P}(X|Y) &= \frac{\mathbb{P}(Y|X)\mathbb{P}(X|Y)\mathbb{P}(X,Y)}{\mathbb{P}(Y|X)\mathbb{P}(X,Y)} \\ &= \mathbb{P}(X|Y)\end{aligned}$$

(e) A survey of certain Montreal students is done, where 55% of the surveyed students are affiliated with UdeM while the others are affiliated with McGill. A student is drawn randomly from this surveyed group.

i. What is the probability that the student is affiliated with McGill?
Answer.

$$\begin{aligned}P(StudentfromMcGill) &= 1 - P(StudentfromUdeM) \\ &= 1 - 0.55 \\ &= 0.45\end{aligned}$$

ii. Now let's say that this student is bilingual, and you know that 80% of UdeM students are bilingual while 50% of McGill students are. Given this information, what is the probability that this student is affiliated with McGill ?

Answer. Using Bayes theorem

$$\begin{aligned}\mathbb{P}(Mcgill|Bilingual) &= \frac{\mathbb{P}(Bilingual|Mcgill)\mathbb{P}(Mcgill)}{\mathbb{P}(Bilingual)} \\ &= \frac{0.50 \times 0.45}{0.50 \times 0.45 \ + \ 0.8 \times 0.55} \\ &= \frac{0.225}{0.665} \\ &= 0.338\end{aligned}$$

2. **Bag of words and single topic model** [10 points] We consider a classification problem where we want to predict the topic of a document from a given corpus (collection of documents). The topic of each

1.6 e2 1 / 1

✓ **- 0 pts** **Correct**

- 1 pts did not answer the question

- 1 pts wrong result

- 0.5 pts calculation error

- 0.25 pts did not make the calculation

- 0.25 pts rounding error

- 0 pts no answer or no page selected

document can either be *sports* or *politics*. 2/3 of the documents in the corpus are about *sports* and 1/3 are about *politics*.

We will use a very simple model where we ignore the order of the words appearing in a document and we assume that words in a document are independent from one another given the topic of the document.

In addition, we will use very simple statistics of each document as features: the probabilities that a word chosen randomly in the document is either "goal", "kick", "congress", "vote", or any another word (denoted by *other*). We will call these five categories the vocabulary or dictionary for the documents: $V = \{"goal", "kick", "congress", "vote", other\}$.

Consider the following distributions over words in the vocabulary given a particular topic:

	$\mathbb{P}(\text{word} \mid \text{topic} = sports)$	$\mathbb{P}(\text{word} \mid \text{topic} = politics)$
word = "*goal*"	1/100	7/1000
word = "*kick*"	1/200	3/1000
word = "*congress*"	0	1/50
word = "*vote*"	5/1000	1/100
word = *other*	980/1000	960/1000

Table 1:

This table tells us for example that the probability that a word chosen at random in a document is "vote" is only 5/1000 if the topic of the document is *sport*, but it is 1/100 if the topic is *politics*.

(a) What is the probability that a random word in a document is "goal" given that the topic is *politics*?
Answer.
The conditional probability is given in the question and can be inferred directly from there.

$$\mathbb{P}(w = goal | t = politics) = \frac{7}{1000}$$

(b) In expectation, how many times will the word "goal" appear in a document containing 200 words whose topic is *sports*?

2.1 a **2 / 2**

✓ **- 0 pts** **Correct**

- 2 pts Wrong

- 2 pts No selected pages

document can either be *sports* or *politics*. 2/3 of the documents in the corpus are about *sports* and 1/3 are about *politics*.

We will use a very simple model where we ignore the order of the words appearing in a document and we assume that words in a document are independent from one another given the topic of the document.

In addition, we will use very simple statistics of each document as features: the probabilities that a word chosen randomly in the document is either "goal", "kick", "congress", "vote", or any another word (denoted by *other*). We will call these five categories the vocabulary or dictionary for the documents: $V = \{"goal", "kick", "congress", "vote", other\}$.

Consider the following distributions over words in the vocabulary given a particular topic:

	$\mathbb{P}(\text{word} \mid \text{topic} = sports)$	$\mathbb{P}(\text{word} \mid \text{topic} = politics)$
word = "*goal*"	1/100	7/1000
word = "*kick*"	1/200	3/1000
word = "*congress*"	0	1/50
word = "*vote*"	5/1000	1/100
word = *other*	980/1000	960/1000

Table 1:

This table tells us for example that the probability that a word chosen at random in a document is "vote" is only 5/1000 if the topic of the document is *sport*, but it is 1/100 if the topic is *politics*.

(a) What is the probability that a random word in a document is "goal" given that the topic is *politics*?
Answer.
The conditional probability is given in the question and can be inferred directly from there.

$$\mathbb{P}(w = goal | t = politics) = \frac{7}{1000}$$

(b) In expectation, how many times will the word "goal" appear in a document containing 200 words whose topic is *sports*?

Answer. Since the probability of the word goal in 100 words is 1/100, therefore the expectation of the word goal in 200 words is

$$200 \times 1/100 = 2$$

(c) We draw randomly a document from the corpus. What is the probability that a random word of this document is "goal"?
Answer.

$$\begin{aligned}\mathbb{P}(goal) &= \mathbb{P}(goal|sports)\mathbb{P}(sports) + \mathbb{P}(goal|politics)\mathbb{P}(politics) \\ &= \frac{1}{100} \times \frac{2}{3} + \frac{7}{1000} \times \frac{1}{3} \\ &= \frac{9}{1000}\end{aligned}$$

(d) Suppose that we draw a random word from a document and this word is "kick". What is the probability that the topic of the document is *sports*?
Answer. Using Bayes theorem

$$\begin{aligned}\mathbb{P}(sports|kick) &= \frac{\mathbb{P}(kick|sports)\mathbb{P}(sports)}{\mathbb{P}(kick)} \\ where, P(kick) &= P(kick|sports)P(sports) + P(kick|politics)P(sports) \\ \mathbb{P}(sports|kick) &= \frac{\frac{1}{200} \times \frac{2}{3}}{\frac{1}{200} \times \frac{2}{3} + \frac{3}{1000} \times \frac{1}{3}} \\ &= \frac{10}{13}\end{aligned}$$

(e) Suppose that we randomly draw two words from a document and the first one is "kick". What is the probability that the second word is "goal"?
Answer.

$$\begin{aligned}\mathbb{P}(w = goal|w_{prev} = kick) = \mathbb{P}(w = goal, t = sports|w_{prev} = kick) \\ + \mathbb{P}(w = goal, t = politics|w_{prev} = kick)\end{aligned} \tag{3}$$

Using P(A,B|C) = P(A|B,C) P(B|C) and P(A|B,C)=P(A|B) because A and C independent on B

2.2 b **2 / 2**

✓ **- 0 pts** **Correct**

- 2 pts Wrong

- 2 pts No pages selected

- 2 pts Answer not on selected page

Answer. Since the probability of the word goal in 100 words is 1/100, therefore the expectation of the word goal in 200 words is

$$200 \times 1/100 = 2$$

(c) We draw randomly a document from the corpus. What is the probability that a random word of this document is "goal"?
Answer.

$$\begin{aligned}\mathbb{P}(goal) &= \mathbb{P}(goal|sports)\mathbb{P}(sports) + \mathbb{P}(goal|politics)\mathbb{P}(politics) \\ &= \frac{1}{100} \times \frac{2}{3} + \frac{7}{1000} \times \frac{1}{3} \\ &= \frac{9}{1000}\end{aligned}$$

(d) Suppose that we draw a random word from a document and this word is "kick". What is the probability that the topic of the document is *sports*?
Answer. Using Bayes theorem

$$\begin{aligned}\mathbb{P}(sports|kick) &= \frac{\mathbb{P}(kick|sports)\mathbb{P}(sports)}{\mathbb{P}(kick)} \\ where, P(kick) &= P(kick|sports)P(sports) + P(kick|politics)P(sports) \\ \mathbb{P}(sports|kick) &= \frac{\frac{1}{200} \times \frac{2}{3}}{\frac{1}{200} \times \frac{2}{3} + \frac{3}{1000} \times \frac{1}{3}} \\ &= \frac{10}{13}\end{aligned}$$

(e) Suppose that we randomly draw two words from a document and the first one is "kick". What is the probability that the second word is "goal"?
Answer.

$$\begin{aligned}\mathbb{P}(w = goal|w_{prev} = kick) = \mathbb{P}(w = goal, t = sports|w_{prev} = kick) \\ + \mathbb{P}(w = goal, t = politics|w_{prev} = kick)\end{aligned} \tag{3}$$

Using P(A,B|C) = P(A|B,C) P(B|C) and P(A|B,C)=P(A|B) because A and C independent on B

2.3 C **2 / 2**

✓ **- 0 pts** **Correct**

- 0.5 pts Reduce to 9/1000

- 2 pts Wrong

- 2 pts No pages selected

Answer. Since the probability of the word goal in 100 words is 1/100, therefore the expectation of the word goal in 200 words is

$$200 \times 1/100 = 2$$

(c) We draw randomly a document from the corpus. What is the probability that a random word of this document is "goal"?
Answer.

$$\begin{aligned}\mathbb{P}(goal) &= \mathbb{P}(goal|sports)\mathbb{P}(sports) + \mathbb{P}(goal|politics)\mathbb{P}(politics) \\ &= \frac{1}{100} \times \frac{2}{3} + \frac{7}{1000} \times \frac{1}{3} \\ &= \frac{9}{1000}\end{aligned}$$

(d) Suppose that we draw a random word from a document and this word is "kick". What is the probability that the topic of the document is *sports*?
Answer. Using Bayes theorem

$$\begin{aligned}\mathbb{P}(sports|kick) &= \frac{\mathbb{P}(kick|sports)\mathbb{P}(sports)}{\mathbb{P}(kick)} \\ where, P(kick) &= P(kick|sports)P(sports) + P(kick|politics)P(sports) \\ \mathbb{P}(sports|kick) &= \frac{\frac{1}{200} \times \frac{2}{3}}{\frac{1}{200} \times \frac{2}{3} + \frac{3}{1000} \times \frac{1}{3}} \\ &= \frac{10}{13}\end{aligned}$$

(e) Suppose that we randomly draw two words from a document and the first one is "kick". What is the probability that the second word is "goal"?
Answer.

$$\begin{aligned}\mathbb{P}(w = goal|w_{prev} = kick) = \mathbb{P}(w = goal, t = sports|w_{prev} = kick) \\ + \mathbb{P}(w = goal, t = politics|w_{prev} = kick)\end{aligned} \tag{3}$$

Using P(A,B|C) = P(A|B,C) P(B|C) and P(A|B,C)=P(A|B) because A and C independent on B

2.4 d **2 / 2**

✓ **- 0 pts** **Correct**

- 2 pts Wrong

- 1 pts Reduce to 10/13

- 2 pts Answer not on selected page

- 0.5 pts Bad rounding (should be 0.77)

- 2 pts Write final answer not equation

Answer. Since the probability of the word goal in 100 words is 1/100, therefore the expectation of the word goal in 200 words is

$$200 \times 1/100 = 2$$

(c) We draw randomly a document from the corpus. What is the probability that a random word of this document is "goal"?
Answer.

$$\begin{aligned}\mathbb{P}(goal) &= \mathbb{P}(goal|sports)\mathbb{P}(sports) + \mathbb{P}(goal|politics)\mathbb{P}(politics) \\ &= \frac{1}{100} \times \frac{2}{3} + \frac{7}{1000} \times \frac{1}{3} \\ &= \frac{9}{1000}\end{aligned}$$

(d) Suppose that we draw a random word from a document and this word is "kick". What is the probability that the topic of the document is *sports*?
Answer. Using Bayes theorem

$$\begin{aligned}\mathbb{P}(sports|kick) &= \frac{\mathbb{P}(kick|sports)\mathbb{P}(sports)}{\mathbb{P}(kick)} \\ where, P(kick) &= P(kick|sports)P(sports) + P(kick|politics)P(sports) \\ \mathbb{P}(sports|kick) &= \frac{\frac{1}{200} \times \frac{2}{3}}{\frac{1}{200} \times \frac{2}{3} + \frac{3}{1000} \times \frac{1}{3}} \\ &= \frac{10}{13}\end{aligned}$$

(e) Suppose that we randomly draw two words from a document and the first one is "kick". What is the probability that the second word is "goal"?
Answer.

$$\begin{aligned}\mathbb{P}(w = goal|w_{prev} = kick) = \mathbb{P}(w = goal, t = sports|w_{prev} = kick) \\ + \mathbb{P}(w = goal, t = politics|w_{prev} = kick)\end{aligned} \tag{3}$$

Using P(A,B|C) = P(A|B,C) P(B|C) and P(A|B,C)=P(A|B) because A and C independent on B

$$\mathbb{P}(w = goal, t = sports | w_{prev} = kick) = P(goal|sports, kick)P(sports|kick)$$
$$= 1/100 \times 10/13 \tag{4}$$

$$\mathbb{P}(w = goal, t = politics | w_{prev} = kick) = P(goal|politics, kick)P(politics|kick)$$
$$= 7/1000 \times 1 - 10/13 \tag{5}$$

Using 4 and 5 in 3

$$\begin{aligned}\mathbb{P}(w = goal | w_{prev} = kick) &= 1/100 \times 10/13 + 7/1000 \times 3/13 \\ &= 121/13000 \\ &= 0.0093\end{aligned}$$

(f) Going back to learning, suppose that you do not know the conditional probabilities given a topic or the probability of each topic (i.e. you don't have access to the information in table 1 or the topic distribution), but you have a dataset of N documents where each document is labeled with one of the topics *sports* and *politics*. How would you estimate the conditional probabilities (e.g., $\mathbb{P}$(word = "*goal*" | topic = *politics*)) and topic probabilities (e.g., $\mathbb{P}$(topic = *politics*)) from this dataset?

Answer.

Since we have the documents labelled with one of the topic, it becomes easy to estimate the topic probabilities.

$$\mathbb{P}(topic = politics) = \frac{No.\ of\ documents\ labelled\ as\ politics}{N}$$
$$\mathbb{P}(topic = sports) = \frac{No.\ of\ documents\ labelled\ as\ sports}{N}$$

The conditional probabilities of the word given a topic can also be calculated by finding the term frequency of the word and dividing it by the total number of words in that document. i.e. fraction of times the word w_i appears among all words in documents of topic t_j.

We can create a mega document by concatenating all the docs from a particular topic j and then finding the frequency of the

2.5 e **2 / 2**

✓ **- 0 pts** **Correct**

- 2 pts Wrong

- 1 pts Reduce to 121/13000

- 2 pts Write answer not equation

- 2 pts Answer not on selected page

- 1 pts Bad rounding

$$\mathbb{P}(w = goal, t = sports | w_{prev} = kick) = P(goal|sports, kick)P(sports|kick)$$
$$= 1/100 \times 10/13 \tag{4}$$

$$\mathbb{P}(w = goal, t = politics | w_{prev} = kick) = P(goal|politics, kick)P(politics|kick)$$
$$= 7/1000 \times 1 - 10/13 \tag{5}$$

Using 4 and 5 in 3

$$\begin{aligned}\mathbb{P}(w = goal | w_{prev} = kick) &= 1/100 \times 10/13 + 7/1000 \times 3/13 \\ &= 121/13000 \\ &= 0.0093\end{aligned}$$

(f) Going back to learning, suppose that you do not know the conditional probabilities given a topic or the probability of each topic (i.e. you don't have access to the information in table 1 or the topic distribution), but you have a dataset of N documents where each document is labeled with one of the topics *sports* and *politics*. How would you estimate the conditional probabilities (e.g., $\mathbb{P}$(word = "*goal*" | topic = *politics*)) and topic probabilities (e.g., $\mathbb{P}$(topic = *politics*)) from this dataset?

Answer.

Since we have the documents labelled with one of the topic, it becomes easy to estimate the topic probabilities.

$$\mathbb{P}(topic = politics) = \frac{No.\ of\ documents\ labelled\ as\ politics}{N}$$
$$\mathbb{P}(topic = sports) = \frac{No.\ of\ documents\ labelled\ as\ sports}{N}$$

The conditional probabilities of the word given a topic can also be calculated by finding the term frequency of the word and dividing it by the total number of words in that document. i.e. fraction of times the word w_i appears among all words in documents of topic t_j.

We can create a mega document by concatenating all the docs from a particular topic j and then finding the frequency of the

word w. This word w can come from a self-made vocabulary or it can be done on all the words we see in the document.

$$\mathbb{P}(w_i|t_j) = \frac{count(w_i, t_j)}{\Sigma\, count(w, c_j)}$$

Sometimes, we do see that a word does not appear in one of the topic, therefore the probability for that will become 0 which is not the correct representation, so we can use Laplace Smoothing by adding 1 in both numerator and denominator to correct it.

3. **Maximum likelihood estimation** [5 points]

Let $x \in \mathbb{R}$ be uniformly distributed in the interval $[0, \theta]$ where θ is a parameter. That is, the pdf of x is given by

$$f_\theta(x) = \begin{cases} 1/\theta & \text{if } 0 \leq \mathbf{x} \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

Suppose that n samples $D = \{x_1, \ldots, x_n\}$ are drawn independently according to $f_\theta(x)$.

(a) Let $f_\theta(x_1, x_2, \ldots, x_n)$ denote the joint pdf of n independent and identically distributed (i.i.d.) samples drawn according to $f_\theta(x)$. Express $f_\theta(x_1, x_2, \ldots, x_n)$ as a function of $f_\theta(x_1), f_\theta(x_2), \ldots, f_\theta(x_n)$
Answer.
Since the n observations are independent and identically distributed (i.i.d), we can give the joint distribution as..

$$\begin{aligned} f_\theta(x_1, x_2, \ldots, x_n) &= f(x_1, x_2, ..., x_n|\theta) \\ &= f(x_1|\theta) f(x_2|\theta), ..., f(x_n|\theta) \\ &= f_\theta(x_1) f_\theta(x_2), ..., f_\theta(x_n) \end{aligned}$$

(b) We define the maximum likelihood estimate by the value of θ which maximizes the likelihood of having generated the dataset D from the distribution $f_\theta(x)$. Formally,

$$\theta_{MLE} = \arg\max_{\theta \in \mathbb{R}} f_\theta(x_1, x_2, \ldots, x_n),$$

2.6 f **2 / 2**

✓ **- 0 pts** **Correct**

- 1 pts No document topic probabilities

- 1 pts Wrong conditional probabilities

- 2 pts Wrong

- 2 pts Answer not on selected page

- 0.5 pts No need for sampling

- 0.5 pts Be more specific

word w. This word w can come from a self-made vocabulary or it can be done on all the words we see in the document.

$$\mathbb{P}(w_i|t_j) = \frac{count(w_i, t_j)}{\Sigma\, count(w, c_j)}$$

Sometimes, we do see that a word does not appear in one of the topic, therefore the probability for that will become 0 which is not the correct representation, so we can use Laplace Smoothing by adding 1 in both numerator and denominator to correct it.

3. **Maximum likelihood estimation** [5 points]

Let $x \in \mathbb{R}$ be uniformly distributed in the interval $[0, \theta]$ where θ is a parameter. That is, the pdf of x is given by

$$f_\theta(x) = \begin{cases} 1/\theta & \text{if } 0 \leq \mathbf{x} \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

Suppose that n samples $D = \{x_1, \ldots, x_n\}$ are drawn independently according to $f_\theta(x)$.

(a) Let $f_\theta(x_1, x_2, \ldots, x_n)$ denote the joint pdf of n independent and identically distributed (i.i.d.) samples drawn according to $f_\theta(x)$. Express $f_\theta(x_1, x_2, \ldots, x_n)$ as a function of $f_\theta(x_1), f_\theta(x_2), \ldots, f_\theta(x_n)$
Answer.
Since the n observations are independent and identically distributed (i.i.d), we can give the joint distribution as..

$$\begin{aligned} f_\theta(x_1, x_2, \ldots, x_n) &= f(x_1, x_2, ..., x_n|\theta) \\ &= f(x_1|\theta) f(x_2|\theta), ..., f(x_n|\theta) \\ &= f_\theta(x_1) f_\theta(x_2), ..., f_\theta(x_n) \end{aligned}$$

(b) We define the maximum likelihood estimate by the value of θ which maximizes the likelihood of having generated the dataset D from the distribution $f_\theta(x)$. Formally,

$$\theta_{MLE} = \arg\max_{\theta \in \mathbb{R}} f_\theta(x_1, x_2, \ldots, x_n),$$

3.1 a **1 / 1**

✓ **- 0 pts** **Correct**

- 0.5 pts Definition of distribution not rigorous enough

- 1 pts Did not express in terms of f(xi)'s

- 1 pts no answer or no page selected

word w. This word w can come from a self-made vocabulary or it can be done on all the words we see in the document.

$$\mathbb{P}(w_i|t_j) = \frac{count(w_i, t_j)}{\Sigma\, count(w, c_j)}$$

Sometimes, we do see that a word does not appear in one of the topic, therefore the probability for that will become 0 which is not the correct representation, so we can use Laplace Smoothing by adding 1 in both numerator and denominator to correct it.

3. **Maximum likelihood estimation** [5 points]

Let $x \in \mathbb{R}$ be uniformly distributed in the interval $[0, \theta]$ where θ is a parameter. That is, the pdf of x is given by

$$f_\theta(x) = \begin{cases} 1/\theta & \text{if } 0 \leq \mathbf{x} \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

Suppose that n samples $D = \{x_1, \ldots, x_n\}$ are drawn independently according to $f_\theta(x)$.

(a) Let $f_\theta(x_1, x_2, \ldots, x_n)$ denote the joint pdf of n independent and identically distributed (i.i.d.) samples drawn according to $f_\theta(x)$. Express $f_\theta(x_1, x_2, \ldots, x_n)$ as a function of $f_\theta(x_1), f_\theta(x_2), \ldots, f_\theta(x_n)$
Answer.
Since the n observations are independent and identically distributed (i.i.d), we can give the joint distribution as..

$$\begin{aligned} f_\theta(x_1, x_2, \ldots, x_n) &= f(x_1, x_2, ..., x_n|\theta) \\ &= f(x_1|\theta) f(x_2|\theta), ..., f(x_n|\theta) \\ &= f_\theta(x_1) f_\theta(x_2), ..., f_\theta(x_n) \end{aligned}$$

(b) We define the maximum likelihood estimate by the value of θ which maximizes the likelihood of having generated the dataset D from the distribution $f_\theta(x)$. Formally,

$$\theta_{MLE} = \arg\max_{\theta \in \mathbb{R}} f_\theta(x_1, x_2, \ldots, x_n),$$

Show that the maximum likelihood estimate of θ is $max(x_1, \ldots, x_n)$

Answer.

$$f_\theta(x) = \begin{cases} 1/\theta & \text{if } 0 \leq \mathbf{x} \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} 1/\theta$$

To maximize the likelihood function we must minimize the value of θ as they have an inverse relation.
Since $f_\theta(x) = 0 \, for \, \theta \leq max \, x_n$, the θ has to be at least max x_n

Therefore the maximum likelihood estimate of θ is $\max(x_1, ..., x_n)$

4. **Maximum likelihood estimation 2** [10 points]

Consider the following probability density function:

$$f_\theta(x) = 2\theta x e^{-\theta x^2}$$

where θ is a parameter and x is positive real number.

Using the same notation as in exercise 3, compute the maximum likelihood estimate of θ.
(hint: you may simplify computations by proving that the maximizer of $f_\theta(x_1, x_2, \ldots, x_n)$ is also the maximizer of $log[f_\theta(x_1, x_2, \ldots, x_n)]$)

Answer.
Assuming the observations are independent and identically distributed and taking log likelihood for simplification.

3.2 b **3.5 / 4**

✓ **- 0 pts** **Correct**

✓ **- 0.5 pts** **Definition of the likelihood/or loglikelihood function not rigorous enough**

- 1.5 pts Did not define/or derive properly the likelihood/ or loglikelihood function

- 1.5 pts Insufficient proof: did not explain/ or incorrectly explained the behavior of the likelihood or the loglikelihood in terms of theta

- 1 pts Insufficient proof: did not explain/ or incorrectly explained the domain of theta in likelihood

- 0.5 pts The xi's are not necessarily ordered

- 4 pts no answer or no page selected

Show that the maximum likelihood estimate of θ is $max(x_1, \ldots, x_n)$

Answer.

$$f_\theta(x) = \begin{cases} 1/\theta & \text{if } 0 \leq \mathbf{x} \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} 1/\theta$$

To maximize the likelihood function we must minimize the value of θ as they have an inverse relation.
Since $f_\theta(x) = 0 \, for \, \theta \leq max \, x_n$, the θ has to be at least max x_n

Therefore the maximum likelihood estimate of θ is $\max(x_1, ..., x_n)$

4. **Maximum likelihood estimation 2** [10 points]

Consider the following probability density function:

$$f_\theta(x) = 2\theta x e^{-\theta x^2}$$

where θ is a parameter and x is positive real number.

Using the same notation as in exercise 3, compute the maximum likelihood estimate of θ.
(hint: you may simplify computations by proving that the maximizer of $f_\theta(x_1, x_2, \ldots, x_n)$ is also the maximizer of $log[f_\theta(x_1, x_2, \ldots, x_n)]$)

Answer.
Assuming the observations are independent and identically distributed and taking log likelihood for simplification.

Taking first derivative of $LL(\theta; x)$ function w.r.t θ and equate it to 0

$$f_\theta(x) = 2\theta x e^{-\theta x^2}$$

$$L(\theta; x) = f_\theta(x_1, x_2, x_3..x_n|\theta)$$

$$= f_\theta(x_1|\theta) f_\theta(x_2|\theta) f_\theta(x_3|\theta)...f_\theta(x_n|\theta)$$

$$LL(\theta; x) = log(f_\theta(x_1))log(f_\theta(x_2))log(f_\theta(x_3))...log(f_\theta(x_n))$$

$$= log(2\theta x_1 e^{-\theta x_1^2})log(2\theta x_2 e^{-\theta x_2^2})log(2\theta x_3 e^{-\theta x_3^2})...log(2\theta x_n e^{-\theta x_n^2})$$

$$= log(2\theta x_1) + log(e^{-\theta x_1^2}) + log(2\theta x_2) + log(e^{-\theta x_n^2})...log(2\theta x_1) + log(e^{-\theta x_n^2})$$

$$\frac{\partial LL(\theta; x)}{\partial \theta} = \frac{2x_1}{2\theta x_1} - x_1^2 + \frac{2x_2}{2\theta x_2} - x_2^2 + ... \frac{2x_n}{2\theta x_n} - x_n^2 = 0$$

$$= \frac{n}{\theta} - (x_1^2 + x_2^2 + ... + x_n^2) = 0$$

$$\theta = \frac{n}{x_1^2 + x_2^2 + ... + x_n^2}$$

5. **k-nearest neighbors** [10 points]

   Let $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n independent labelled samples drawn using the following sampling process:

   - the label of each $\mathbf{x}_i$ is drawn randomly with 50% probability for each of the two classes
   - x_i is drawn uniformly in S^+ if its label is positive, and uniformly in S^- otherwise

   Where S^+ and S^- are two **unit** hyperspheres whose centers are 10 units apart.

4 Maximum likelihood estimation 2 7 / 10

- **0 pts** Correct
- **2 pts** Calculation mistake
- **5 pts** Did not derive likelihood properly or did not derive likelihood for n samples

✓ **- 2 pts** **Did not prove stationary point was a maxima**

✓ **- 1 pts** **Did not mention log(x) is a monotonic/increasing function to prove the use of log**

- **1 pts** Did not simplify expression enough
- **10 pts** no answer or no page selected

(a) Show that if k is odd the average probability of error of the k-NN classifier is given by

$$P_n(e) = \frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j}.$$

Answer.
There is no overlap in the positive and negative labels as the two units hyperspheres are 10 units apart. To predict a class label for a new test case, we need more than (k+1)/2 classes belonging from one of the label.
Also, P(positive) = P(negative) = $\frac{1}{2}$

$P_n(e) = P$(true label is S^+ when more frequent is S^-)
$+ P$(true label is S^- when more frequent is S^+)

$= 2P$(true label is S^+, less labelled are S^+ and more frequent is S^-)

$= 2P(\mathrm{S}^+)$ P(S^+ labelled less than k-1/2 and S^- labelled more than half of k)

= (2)(1/2)(choosing 0,..,j S^+ values)(prob of S^+ values for 0,..,j)
(prob of S^- values for j+1,..,n)

$= (1)(\binom{n}{0}\binom{n}{1}..\binom{n}{k-1/2})(\frac{1}{2}\frac{1}{2}\frac{1}{2}$..j times)$(\frac{1}{2}\frac{1}{2}\frac{1}{2}$..n-j times)

$$= \sum_{j=0}^{(k-1)/2} (\binom{n}{j})(\frac{1}{2})^j(\frac{1}{2})^{n-j}$$

$$= \frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j}$$

5.1 a **1 / 4**

- **0 pts** Correct
- **4 pts** no answer
- **1.5 pts** reasoning error / answer not detailed enough
- ✓ **- 3 pts** **only traces of right answer**
- **4 pts** wrong
- **0.5 pts** you haven't selected all pages corresponding to your answer
- **0 pts** ALMOST ILLEGIBLE !
- **0 pts** DUP 1?
- **2 pts** k can't be 2n+1
- **0 pts** DUP 2?
- **4 pts** no pages selected

(b) Show that in this case the single-nearest neighbor classifier ($k = 1$) has a lower error rate than the k-NN classifier for $k > 1$.

Answer.

$$
\begin{aligned}
for\, k = 1, P(e) &= \frac{1}{2^n} \sum_{j=0}^{(0)} \binom{n}{0} \\
&= \frac{1}{2^n} \\
for\, k > 1, P(e) &= \frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j} \\
&= \frac{1}{2^n} \binom{n}{0} \binom{n}{1} \binom{n}{2} \ldots \binom{n}{(k-1)/2}
\end{aligned}
$$

Since for k > 1, the positive values will be added to $\frac{1}{2^n}$

Therefore, P(e) for k = 1 < P(e) for k >1

(c) If k is allowed to increase with n but is restricted by $k \leq a\sqrt{n}$ (for some constant a), show that $P_n(e) \to 0$ as $n \to \infty$.

Answer. Finding an upper bound on the binomial distribution to prove that P(e) $\to$ 0 as n $\to \infty$

As per tail bounds inequality for a binomial distribution for a function F

$$F(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}, F(k, n, p) \leq \exp(-2 \frac{(np-k)^2}{n} \quad (6)$$

We can use 6 in our case

$$\frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j} \leq exp - 2\frac{(np-k)^2}{n}$$

For the function $e^{-x} \to 0$, x $\to \infty$

5.2 b **0 / 2**

- 0 pts Correct

✓ **- 2 pts wrong**

- 1 pts Starting from the desired result and going through implications to something that's true is not how proofs work in math. If you want to do this, you need to replace your implications by a sentence like "it suffices to prove that..." or use equivalences if possible. If you haven't written anything between equations in different lines, I assume it's an implication !

- 1.5 pts you didn't generalize to all k

- 2 pts wrong induction

- 2 pts no answer

- 0.5 pts "at least n" is wrong

- 0.5 pts we don't *have to* prove Pn(e) is an increasing fct wrt k

- 1 pts k can't be 2

- 0.5 pts you didn't conclude...

- 2 pts no pages selected

(b) Show that in this case the single-nearest neighbor classifier ($k = 1$) has a lower error rate than the k-NN classifier for $k > 1$.

Answer.

$$
\begin{aligned}
for\, k = 1, P(e) &= \frac{1}{2^n} \sum_{j=0}^{(0)} \binom{n}{0} \\
&= \frac{1}{2^n} \\
for\, k > 1, P(e) &= \frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j} \\
&= \frac{1}{2^n} \binom{n}{0} \binom{n}{1} \binom{n}{2} \ldots \binom{n}{(k-1)/2}
\end{aligned}
$$

Since for k > 1, the positive values will be added to $\frac{1}{2^n}$

Therefore, P(e) for k = 1 < P(e) for k >1

(c) If k is allowed to increase with n but is restricted by $k \leq a\sqrt{n}$ (for some constant a), show that $P_n(e) \to 0$ as $n \to \infty$.

Answer. Finding an upper bound on the binomial distribution to prove that P(e) $\to$ 0 as n $\to \infty$

As per tail bounds inequality for a binomial distribution for a function F

$$F(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}, F(k, n, p) \leq \exp(-2 \frac{(np-k)^2}{n} \quad (6)$$

We can use 6 in our case

$$\frac{1}{2^n} \sum_{j=0}^{(k-1)/2} \binom{n}{j} \leq exp - 2\frac{(np-k)^2}{n}$$

For the function $e^{-x} \to 0$, x $\to \infty$

$$\lim_{n\to\infty} 2\frac{(n(1/2)-(k-1)/2)^2}{n}$$

$$\lim_{n\to\infty} 2\frac{(\frac{n-(k-1)}{2})^2}{n}$$

$$\lim_{n\to\infty} \frac{n^2+k^2+1-2k-2nk+2n}{2n}$$

Since k $\leq$ a$\sqrt{n}$

$$\lim_{n\to\infty} n/2+1/2n-a/\sqrt{n}+\ constant$$

Clearly this is going to ∞, *when*, $n \to \infty$

Hence, $P(e) \to 0$, *when* $n \to \infty$

6. **Gaussian Mixture** [10 points]

Let $\mu_1, \mu_2 \in \mathbb{R}^2$, and let Σ_1, Σ_2 be two 2x2 positive definite matrices (i.e. symmetric with positive eigenvalues).
We now introduce the two following pdf over $\mathbb{R}^2$:

$$f_{\mu_1,\Sigma_1}(\mathbf{x}) = \frac{1}{2\pi\sqrt{det(\Sigma_1)}} e^{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)}$$

$$f_{\mu_2,\Sigma_2}(\mathbf{x}) = \frac{1}{2\pi\sqrt{det(\Sigma_2)}} e^{-\frac{1}{2}(\mathbf{x}-\mu_2)^T\Sigma_2^{-1}(\mathbf{x}-\mu_2)}$$

These pdf correspond to the multivariate Gaussian distribution of mean μ_1 and covariance Σ_1, denoted $\mathcal{N}_2(\mu_1, \Sigma_1)$, and the multivariate Gaussian distribution of mean μ_2 and covariance Σ_2, denoted $\mathcal{N}_2(\mu_2, \Sigma_2)$.

We now toss a balanced coin Y, and draw a random variable X in $\mathbb{R}^2$, following this process : if the coin lands on tails ($Y = 0$) we draw X from $\mathcal{N}_2(\mu_1, \Sigma_1)$, and if the coin lands on heads ($Y = 1$) we draw X from $\mathcal{N}_2(\mu_2, \Sigma_2)$.

5.3 C **3 / 4**

- **0 pts** Correct
- **4 pts** wrong
- **0.5 pts** what about k=1, 3, 5 ?
- **0.5 pts** a sqrt(n) is not necessarily an integer and/or odd

✓ **- 1 pts error, (check the box)**

- **0.5 pts** you can't talk about lim P_n(e) or lim 1/2^n sum(...) before proving their existence
- **0.5 pts** a limit can't be equal to something that depends on n
- **4 pts** no pages selected
- **4 pts** no answer
- **3 pts** traces of answer only
- **0.5 pts** correct but needs more rigor
- **2 pts** errors (check the boxes)
- **1 pts** error (check the comment)
- **0.5 pts** a few mistakes in the induction, but mostly ok
- **1.5 pts** error, check the comment
- **0 pts** DUP 1?
- **0 pts** DUP 2?
- **0.5 pts** n/2 not necessarily integer
- **0.5 pts** error (check comment)
- **0.5 pts** limits are not "equivalent to each other"
- **2 pts** you just showed that P_n(e) < 1, while your proof strategy can be quickly adapted to show that P_n(e) goes to 0
- **1 pts** it doesn't matter that the denominator is O(2^n), what matters is that it's Omega(2^n)
- **1 pts** You just proved that a the cdf of a certain binomial RV evaluated at a certain point converges to 0. You didn't make the link with P_n(e)
- **0.5 pts** you didn't select all the pages corresponding to your answer

1 lacks a sum here

Calculate $\mathbb{P}(Y = 0|X = \mathbf{x})$, the probability that the coin landed on tails given $X = \mathbf{x} \in \mathbb{R}^2$, as a function of μ_1, μ_2, Σ_1, Σ_2, and $\mathbf{x}$. Show all the steps of the derivation.

Answer. Using Bayes Theorem

$$\mathbb{P}(Y = 0|X = x) = \frac{\mathbb{P}(X = x|Y = 0)\mathbb{P}(Y = 0)}{\mathbb{P}(X)}$$

$$= \frac{\mathbb{P}(X = x|Y = 0)\mathbb{P}(Y = 0)}{\mathbb{P}(X = x|Y = 0)\mathbb{P}(Y = 0) \; + \; \mathbb{P}(X = x|Y = 1)\mathbb{P}(Y = 1)}$$

$$= \frac{\frac{1}{2\pi\sqrt{det(\Sigma_1)}} e^{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)}}{\frac{1}{2\pi\sqrt{det(\Sigma_1)}} e^{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)} \; + \; \frac{1}{2\pi\sqrt{det(\Sigma_2)}} e^{-\frac{1}{2}(\mathbf{x}-\mu_2)^T\Sigma_2^{-1}(\mathbf{x}-\mu_2)}}$$

$$= \frac{1}{1 \; + \; \frac{\sqrt{det(\Sigma_1)}}{\sqrt{det(\Sigma_2)}} e^{-\frac{1}{2}[(\mathbf{x}-\mu_2)^T\Sigma_2^{-1}(\mathbf{x}-\mu_2) - (\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)]}}$$

6 Gaussian Mixture 10 / 10

✓ **- 0 pts Correct**

- 1 pts result is not simplified

- 2 pts calculation error : sigma1 and sigma2 exchanged

- 2 pts calculation error : additional or missing term

- 10 pts wrong reasoning

- 2 pts calculation error

- 2 pts notation error / lack of rigor

- 7.5 pts serious mathematical error

- 2 pts calculation error : wrong sign

- 7.5 pts serious calculation error

- 2 pts result is not simplified at all

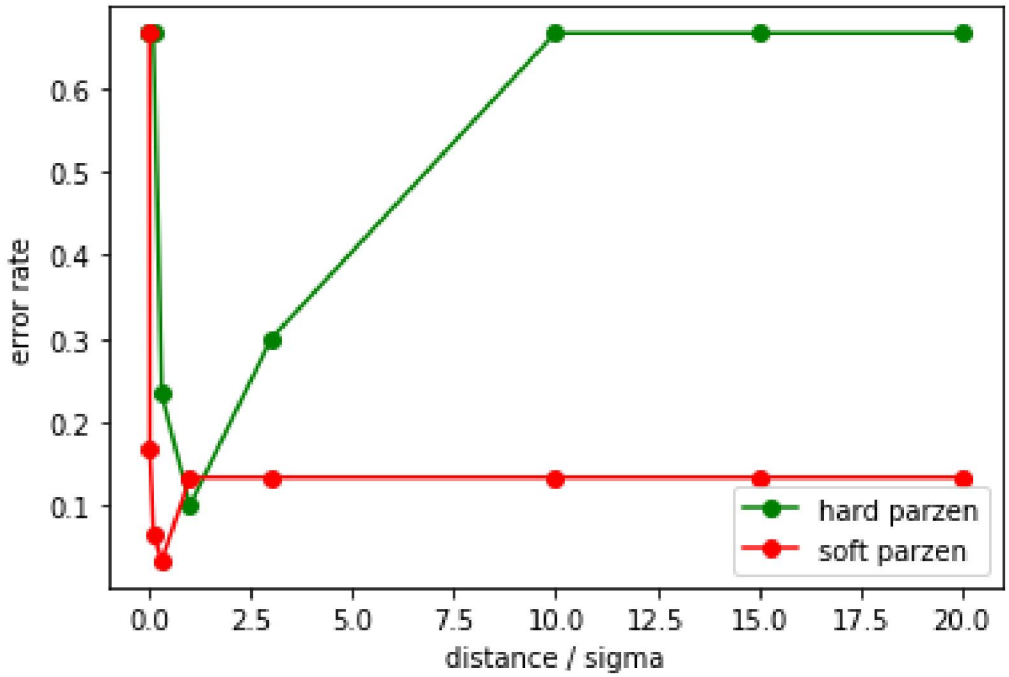
- 5 pts unnecessary reasoning error / confusion

- 10 pts no answer or no page selected

- 1 pts illegible

- 10 pts no justification

- 5 pts unsufficient justification



Answer. As evident, soft parzen has a lower error rate when compared to hard parzen because it takes a weighted vote for all the training points.
Soft Parzen: The bandwidth or the sigma exhibits a strong influence on the resulting estimate. A large width will over smooth the density and mask the structure in the data whereas a small bandwidth will yield a density estimate that is spiky and very hard to interpret. For σ=0.3 the error is minimized between the estimated density and the true density giving an optimal solution. But once we increase the sigma, the error increases and then becomes constant because the weight of all training points will not change with the further increase in bandwidth/width.

Hard Parzen: Similarily, the radius h plays an important role in evaluating the performance of our method. Our classifier becomes blind to the overall data when the h is too small whereas high value of h will smoothen the decision boundaries and give a more erratic prediction as the classifier tends to get biased towards the dominant class with more training points being added for a vote. On further increase of the radius, the error rate seems to become constant because by then we had already covered all the points in the training data and there are no more points left to change the error rate.

7.1 practical report Q5 7 / 7

✓ **- 0 pts Correct**

- 2.5 pts error in one of the curves

- 5 pts error in both curves

- 2 pts no discussion / comment

- 1 pts incomplete discussion / comment

- 1 pts one of the curves is not totally right

- 7 pts no pages selected

- 7 pts no answer

- 0.5 pts it is literally mentioned in the homework : "in the same plot"

- 0 pts Click here to replace this description.

- 2 pts wrong axes

Graduates 5 pts

Question. 7. Include in your report a discussion on the running time complexity of these two methods. How does it vary for each method when the hyperparameter h or σ changes? Why ?

Answer.

Both the algorithms have a run time complexity of $O(\mathrm{n})$ where n is the number of training data. Since the distance has to be calculated on all the training points at the time of prediction, the value of radius h or bandwidth σ will have no affect on the time complexity.

Although both the methods have the same complexity, but in real time the hard parzen runs quite faster than soft parzen because of less computations.

Graduates 10 pts

Question. 9. Similar to Question 5, compute the validation errors of Hard Parzen classifiers trained on 500 random projections of the training set, for

$$h \in \{0.001, 0.01, 0.1, 0.3, 1.0, 3.0, 10.0, 15.0, 20.0\}$$

The validation errors should be computed on the projected validation set, using the same matrix A. To obtain random projections, you may draw A as 8 independent variables drawn uniformly from a gaussian distribution of mean 0 and variance 1.

You can for example store these validation errors in a 500×9 matrix, with a row for each random projection and a column for each value of h.

Do the same thing for RBF Parzen classifiers, for

$$\sigma \in \{0.001, 0.01, 0.1, 0.3, 1.0, 3.0, 10.0, 15.0, 20.0\}$$

Plot and include in your report in the same graph the average values of the validation errors (over all random projections) for each value of h and σ, along with error bars of length equal to $0.2\times$ the standard deviations.How do your results compare to the previous ones?

7.2 practical report Q7 5 / 5

✓ **- 0 pts Correct**

- 0 pts no variation in h

- 0 pts soft parzen < hard parzen

- 0 pts weird time curves

- 2.5 pts wrong interpretation (soft)

- 0 pts ignored practical running time

- 5 pts false reasoning

- 5 pts absurd results

- 2.5 pts wrong interpretation (hard)

- 5 pts page not selected or no answer

- 2 pts hard parzen complexity is not proportional to neighbors

- 2.5 pts does not explain dependency

- 3 pts interpretation too superficial or does not match obs.

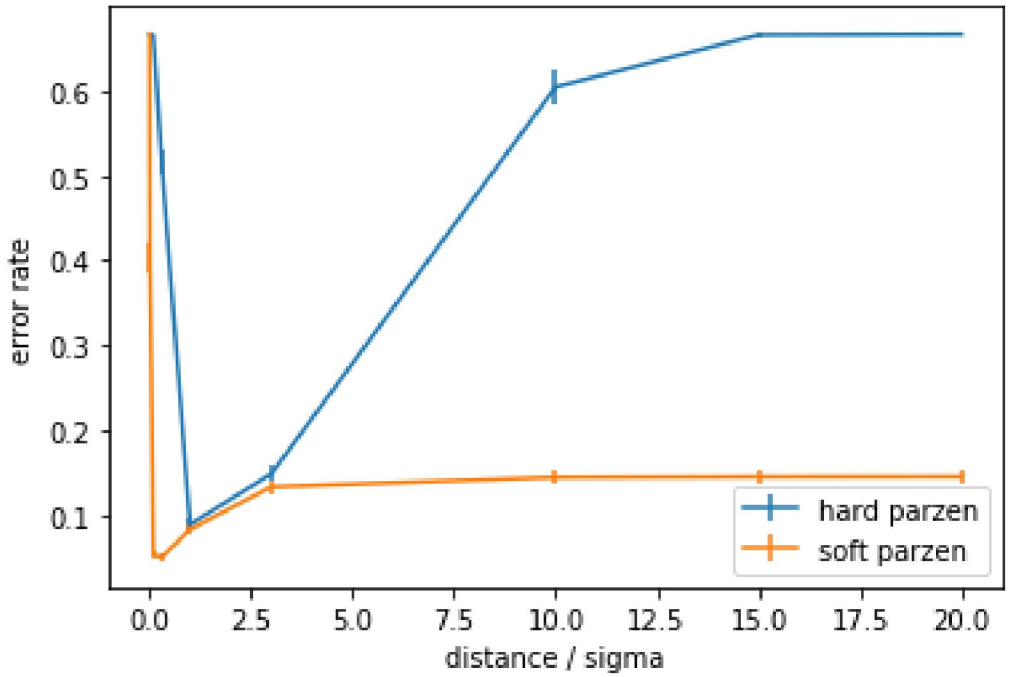
- 5 pts does not answer the question

- 3 pts partial error in interpretation

- 1.5 pts error in theo. complexity calculation

- 2 pts unclear justification

- 1 pts hardly legible



Answer.

The random projection from a gaussian distribution doesnt seem to have any affect on the error rate plot as the results in this graph seems to corroborate the one we had earlier. The error for soft parzen is lower than that of hard parzen. As usual, the hard parzen error rises with increase in radius because it takes too many training data points in consideration for prediction and then becomes constant when all the points are considered. Whereas in the soft parzen, the constant error rate is achieved much in advance because all the training samples are in the same region even if we increase the bandwidth anymore.

7.3 practical report Q9 7.5 / 10

- **0 pts** Correct
- ✓ **- 2.5 pts** **missing/wrong error bars**
- **- 4 pts** one of the curves is wrong
- **- 8 pts** both curves are wrong
- **- 2 pts** no comparison / discussion
- **- 2 pts** one of the curves is not exactly right
- **- 4 pts** both curves are not exactly right
- **- 10 pts** no pages selected
- **- 10 pts** no answer
- **- 2 pts** your plot should be a line plot
- **- 0 pts** DUP 1
- **- 0 pts** DUP 2