SC4000/CZ4041/CE4041: Machine Learning
Lecture 2 Tutorial Question Sets

Question 1: Suppose A, B and C are three variables of binary values (0 or 1). Given the probabilities $P(A = 1, B = 0) = 0.4$, $P(A = 0) = 0.3$, and $P(A = 1, B = 1, C = 1) = 0.1$, compute the following probabilities:

1. $P(B = 1|A = 1)$.
2. $P(C = 0|B = 1, A = 1)$.

Question 2: Suppose that if a person has lung cancer, his/her probability of having gene X is 0.9, and if a person does not have lung cancer, his/her probability of having gene X is 0.2. The probability of a person having lung cancer is 0.01. Now, we know that a patient A has gene X.

1. Use Bayesian decision theory with 0/1 loss to predict whether the patient A has lung cancer or not.
2. Consider that costs of misclassification are different. Assume that the cost for correct decisions is 0, the cost of misclassifying a person who does not have lung cancer to be a patient with lung cancer is 0.007, and the cost of misclassifying a person who has lung cancer to be a healthy person is 1. Please use Bayesian decision theory with the predefined loss to predict whether the patient A has lung cancer or not.

Question 1

Question 1: Suppose A, B and C are three variables of binary values (0 or 1). Given the probabilities $P(A = 1, B = 0) = 0.4$, $P(A = 0) = 0.3$, and $P(A = 1, B = 1, C = 1) = 0.1$, compute the following probabilities:

1. $P(B = 1|A = 1)$.
2. $P(C = 0|B = 1, A = 1)$.

1.1

- $P(B = 1|A = 1) = \frac{P(B=1,A=1)}{P(A=1)}$
- $P(A = 1) = 1 - P(A = 0) = 1 - 0.3 = 0.7$
- $P(A = 1) = 0.7$
  $= P(A = 1, B = 0) + P(A = 1, B = 1)$
  $= 0.4 + P(A = 1, B = 1)$
- $P(A = 1, B = 1) = 0.7 - 0.4 = 0.3$
- Thus, $P(B = 1|A = 1) = \frac{0.3}{0.7} = \frac{3}{7}$

1.2

- $P(C = 0|B = 1, A = 1)$
  $= \frac{P(C=0,B=1,A=1)}{P(B=1,A=1)} = \frac{P(C=0,B=1,A=1)}{0.3}$
- $P(B = 1, A = 1) = 0.3$
  $= P(C = 0, B = 1, A = 1) + P(C = 1, B = 1, A = 1)$
  $= P(C = 0, B = 1, A = 1) + 0.1$
- $P(C = 0, B = 1, A = 1) = 0.3 - 0.1 = 0.2$
- Thus, $P(C = 0|B = 1, A = 1) = \frac{0.2}{0.3} = \frac{2}{3}$

Question 2

Question 2: Suppose that if a person has lung cancer, his/her probability of having gene X is 0.9, and if a person does not have lung cancer, his/her probability of having gene X is 0.2. The probability of a person having lung cancer is 0.01. Now, we know that a patient A has gene X.

1. Use Bayesian decision theory with 0/1 loss to predict whether the patient A has lung cancer or not.
2. Consider that costs of misclassification are different. Assume that the cost for correct decisions is 0, the cost of misclassifying a person who does not have lung cancer to be a patient with lung cancer is 0.007, and the cost of misclassifying a person who has lung cancer to be a healthy person is 1. Please use Bayesian decision theory with the predefined loss to predict whether the patient A has lung cancer or not.

- Let Y be the random variable that represents a person having lung cancer or not (0 or 1)
  - $Y = 0$: a person does not have lung cancer
  - $Y = 1$: a person has lung cancer
- Let Z be the random variable that represents a person having gene X or not (0 or 1)
  - $Z = 0$: a person does not have gene X
  - $Z = 1$: a person has gene X

2.1

- If a person has lung cancer, his/her probability of having gene X is 0.9
  $P(Z = 1|Y = 1) = 0.9$
  ⬇
  $P(Z = 0|Y = 1) = 0.1$
- If a person does not have lung cancer, his/her probability of having gene X is 0.2
  $P(Z = 1|Y = 0) = 0.2$
  ⬇
  $P(Z = 0|Y = 0) = 0.8$
- The probability of a person having lung cancer is 0.01
  $P(Y = 1) = 0.01$
  ⬇
  $P(Y = 0) = 0.99$

- We know that patient A has gene X, i.e., $Z = 1$
- To estimate $P(Y = 0|Z = 1)$ v.s. $P(Y = 1|Z = 1)$
- With the 0/1 loss

Predict $Y = 1$ if $P(Y = 1|Z = 1) > P(Y = 0|Z = 1)$

Predict $Y = 0$ otherwise

OR $P(Z = 1|Y = 1)P(Y = 1) > P(Z = 1|Y = 0)P(Y = 0)$

- We have
  - $P(Z = 1|Y = 1)P(Y = 1) = 0.9 \times 0.01 = 0.009$
  - $P(Z = 1|Y = 0)P(Y = 0) = 0.2 \times 0.99 = 0.198$
  - $P(Z = 1|Y = 0)P(Y = 0) > P(Z = 1|Y = 1)P(Y = 1)$
  ⬇
  - $P(Y = 0|Z = 1) > P(Y = 1|Z = 1)$
- Prediction: patient A does not have lung cancer

2.2

- λ_{ik}: loss (or cost) of action a_i that predicts $Y = i$ while the true label is k
- The cost for correct decisions is 0
  $\lambda_{00} = 0$, and $\lambda_{11} = 0$
- Cost of misclassifying a person not having lung cancer to be a patient with lung cancer is 0.007
  $\lambda_{10} = 0.007$
- Cost of misclassifying a person having lung cancer to be a healthy person is 1
  $\lambda_{01} = 1$

- Take action a^* if $a^* = \arg\min_{a_i} R(a_i|\boldsymbol{x})$, where
  $R(a_i|\boldsymbol{x}) = \sum_{c=0}^{C-1} \lambda_{ic} P(Y = c|\boldsymbol{x})$
- Expected risk of predicting A not having lung cancer:

$$R(a_0|Z = 1) = \lambda_{01} \times P(Y = 1|Z = 1) + \lambda_{00} \times P(Y = 0|Z = 1)$$

$$= 1 \times \frac{P(Z = 1|Y = 1) \times P(Y = 1)}{P(Z = 1)} + 0$$

$$= 1 \times \frac{0.9 \times 0.01}{P(Z = 1)} = \frac{0.009}{P(Z = 1)}$$

- Expected risk of predicting A having lung cancer:

$$R(a_1|Z = 1) = \lambda_{11} \times P(Y = 1|Z = 1) + \lambda_{10} \times P(Y = 0|Z = 1)$$

$$= 0 + 0.007 \times \frac{P(Z = 1|Y = 0) \times P(Y = 0)}{P(Z = 1)}$$

$$= 0.007 \times \frac{0.198}{P(Z = 1)}$$

$$= \frac{0.0014}{P(Z = 1)} < R(a_0|Z = 1) = \frac{0.009}{P(Z = 1)}$$

- Prediction: patient A has lung cancer

Question 1.2

Task: Predict class label for $(A = 1, B = 1, C = 1)$ using NB

Let $P(A = 1, B = 1, C = 1) = K$

$P(+|A = 1, B = 1, C = 1)$

$$= \frac{P(A = 1, B = 1, C = 1|+) \times P(+)}{K}$$
$$= \frac{P(A = 1|+) \times P(B = 1|+) \times P(C = 1|+) \times P(+)}{K}$$
$$= \frac{0.5 \times 0.5 \times 1 \times 0.4}{K}$$
$$= \frac{0.1}{K}$$

$P(-|A = 1, B = 1, C = 1)$

$$= \frac{P(A = 1, B = 1, C = 1|-) \times P(-)}{K}$$
$$= \frac{P(A = 1|-) \times P(B = 1|-) \times P(C = 1|-) \times P(-)}{K}$$
$$= \frac{0.33 \times 0.33 \times 0.33 \times 0.6}{K}$$
$$= \frac{0.0222}{K} < \frac{0.1}{K}$$ $P(+|A = 1, B = 1, C = 1)$

Class label = "+"

SC4000/CZ4041/CE4041: Machine Learning
Lecture 3 Tutorial Question Sets

Question 1: Consider the data set shown in Table 1.

Question 1.1

Estimate conditional probabilities

$$P(x_i = k|y = c) = \frac{|(x_i = k) \wedge (y = c)|}{|y = c|}$$

Table 1: Data set for Question 1.

Record	A	B	C	Class
1	0	0	0	-
2	0	0	1	-
3	0	1	1	+
4	0	1	1	+
5	0	0	1	-
6	1	0	1	+
7	1	1	0	-
8	1	0	0	-
9	1	0	1	+
10	0	1	0	-

$P(A = 1\|-) = 2/6 = 0.33$	$P(A = 0\|-) = 4/6 = 0.67$
$P(B = 1\|-) = 2/6 = 0.33$	$P(B = 0\|-) = 4/6 = 0.67$
$P(C = 1\|-) = 2/6 = 0.33$	$P(C = 0\|-) = 4/6 = 0.67$
$P(A = 1\|+) = 2/4 = 0.5$	$P(A = 0\|+) = 2/4 = 0.5$
$P(B = 1\|+) = 2/4 = 0.5$	$P(B = 0\|+) = 2/4 = 0.5$
$P(C = 1\|+) = 4/4 = 1$	$P(C = 0\|+) = 0/4 = 0$

Q2

Tid	Home Owner	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

1. Estimate the conditional probabilities for $P(A = 1|+)$, $P(B = 1|+)$, $P(C = 1|+)$, $P(A = 1|-)$, $P(B = 1|-)$, and $P(C = 1|-)$.
2. Use the estimate of conditional probabilities given in the previous question to predict the class label for a test example $(A = 1, B = 1, C = 1)$ using the naïve Bayes approach.

Question 2: On Page 28 of the lecture notes "Lecture 3", recalculate the likelihoods using m-estimate. Compare the m-estimate method with the original method shown on Page 25 for estimating probabilities. Which method is better and why?

Home Owner	Marital Status	Taxable Income	Cheat
No	Married	120K	?

$m = 3$

$p = 1/3$ for all discrete features of class **Yes**

$p = 2/3$ for all discrete features of class **No**

P(HomO=Yes|No) = (3+2)/(7+3) = 5/10
P(HomO=No|No) = (4+2)/(7+3) = 6/10
P(HomO=Yes|Yes) = (0+1)/(3+3) = 1/6
P(HomO=No|Yes) = (3+1)/(3+3) = 4/6

P(Status=Single|No) = (2+2)/(7+3) = 4/10
P(Status=Divorced|No)= (1+2)/(7+3) = 3/10
P(Status=Married|No) = (4+2)/(7+3) = 6/10
P(Status=Single|Yes) = (2+1)/(3+3) = 3/6
P(Status=Divorced|Yes) = (1+1)/(3+3) = 2/6
P(Status=Married|Yes) = (0+1)/(3+3) = 1/6

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

P(Class = No) = 7/10
P(Class = Yes) = 3/10

$$\frac{|(x_i = k) \wedge (y = c)| + m \times p}{|y = c| + m}$$

$$\sum_k P(x_i = k|y = c) = 1$$

$$\hat{P}(\text{HomO} = \text{Yes}|\text{No}) = \frac{P(\text{HomO} = \text{Yes}|\text{No})}{P(\text{HomO} = \text{Yes}|\text{No}) + P(\text{HomO} = \text{No}|\text{No})} = \frac{\frac{5}{10}}{\frac{5}{10} + \frac{6}{10}} = \frac{5}{11}$$

$$\hat{P}(\text{HomO} = \text{No}|\text{No}) = \frac{P(\text{HomO} = \text{No}|\text{No})}{P(\text{HomO} = \text{Yes}|\text{No}) + P(\text{HomO} = \text{No}|\text{No})} = \frac{\frac{6}{10}}{\frac{5}{10} + \frac{6}{10}} = \frac{6}{11}$$

8

Home Owner	Marital Status	Taxable Income	Cheat
No	Married	120K	?

$m = 3$

$p = 1/3$ for all discrete features of class **Yes**

$p = 2/3$ for all discrete features of class **No**

P(HomO=Yes|No) = (3+2)/(7+3) = 5/10
P(HomO=No|No) = (4+2)/(7+3) = 6/10
P(HomO=Yes|Yes) = (0+1)/(3+3) = 1/6
P(HomO=No|Yes) = (3+1)/(3+3) = 4/6

P(Status=Single|No) = (2+2)/(7+3) = 4/10
P(Status=Divorced|No)= (1+2)/(7+3) = 3/10
P(Status=Married|No) = (4+2)/(7+3) = 6/10
P(Status=Single|Yes) = (2+1)/(3+3) = 3/6
P(Status=Divorced|Yes) = (1+1)/(3+3) = 2/6
P(Status=Married|Yes) = (0+1)/(3+3) = 1/6

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

P(Class = No) = 7/10
P(Class = Yes) = 3/10

9

Home Owner	Marital Status	Taxable Income	Cheat
No	Married	120K	?

$m = 3$

$p = 1/3$ for all discrete features of class **Yes**

$p = 2/3$ for all discrete features of class **No**

P(HomO=Yes|No) = (5/10)/(5/10+6/10) = 5/11
P(HomO=No|No) = (6/10)/(5/10+6/10) = 6/11
P(HomO=Yes|Yes) = (1/6)/(1/6+4/6) = 1/5
P(HomO=No|Yes) = (4/6)/(1/6+4/6) = 4/5

P(Status=Single|No) = (4/10)/(4/10+3/10+6/10) = 4/13
P(Status=Divorced|No)= (3/10)/(4/10+3/10+6/10) = 3/13
P(Status=Married|No) = (6/10)/(4/10+3/10+6/10) = 6/13
P(Status=Single|Yes) = (2+1)/(3+3) = 3/6
P(Status=Divorced|Yes) = (1+1)/(3+3) = 2/6
P(Status=Married|Yes) = (0+1)/(3+3) = 1/6

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

P(Class = No) = 7/10
P(Class = Yes) = 3/10

10

Home Owner	Marital Status	Taxable Income	Cheat
No	Married	120K	?

$m = 3$

$p = 1/3$ for all discrete features of class **Yes**

$p = 2/3$ for all discrete features of class **No**

P(HomO=Yes|No) = 5/11
P(HomO=No|No) = 6/11
P(HomO=Yes|Yes) = 1/5
P(HomO=No|Yes) = 4/5

P(Status=Single|No) = 4/13
P(Status=Divorced|No) = 3/13
P(Status=Married|No) = 6/13
P(Status=Single|Yes) = 3/6
P(Status=Divorced|Yes) = 2/6
P(Status=Married|Yes) = 1/6

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

P(Class = No) = 7/10
P(Class = Yes) = 3/10

P(**X**|Class=No) = P(HomO=No | Class=No)
× P(Status=Married | Class=No)
× P(Income=120K | Class=No)
= 6/11 × 6/13 × 0.0072 = 0.0018

P(**X**|Class=Yes) = P(HomO=No | Class=Yes)
× P(Status=Married | Class=Yes)
× P(Income=120K | Class=Yes)
= 4/5×1/6×1.2×10^{-9} = 1.6 ×10^{-10}

P(**X**|No) × P(No) > P(**X**|Yes) × P(Yes)
Thus, P(No|**X**) > P(Yes|**X**)

Cheat = No

SC4000/CZ4041/CE4041: Machine Learning
Lecture 4 Tutorial Question Sets

$P(D=1 \mid E=1, a=1, b=1)$

Question 1: Given the Bayesian Belief Network shown in Figure 1 (the same as on the lecture notes), if the person has high blood pressure, but exercises regularly and eats a healthy die, to estimate the probability that the person has heart disease.

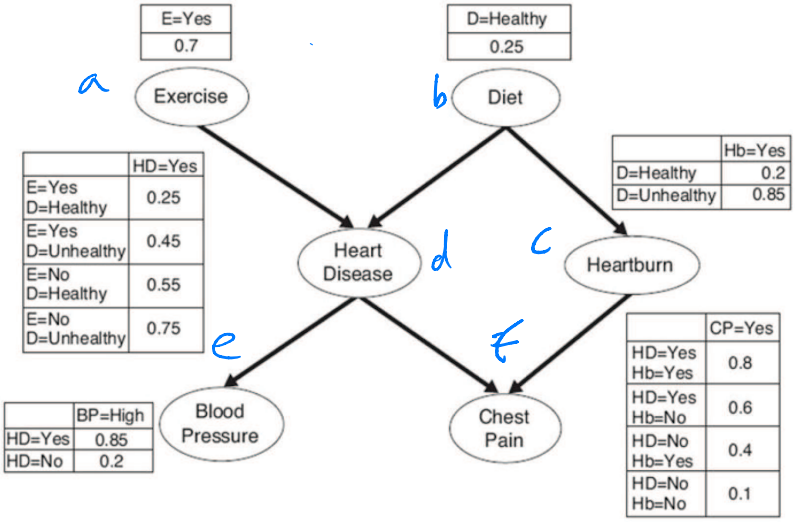


Figure 1: The Bayesian Network for Question 1.

Question 2: Consider the Bayesian Belief Network shown in Figure 2. Given that the outcomes of medical test A and medical test B for a specific patient are positive and negative, respectively. That is, $A =$ Pos and $B =$ Neg. Predict the probability that the patient has high cholesterol.

Pos = 1 Neg = 0

$A = 1$

$B = 0$

$P(C=1 \mid A=1, B=0)$

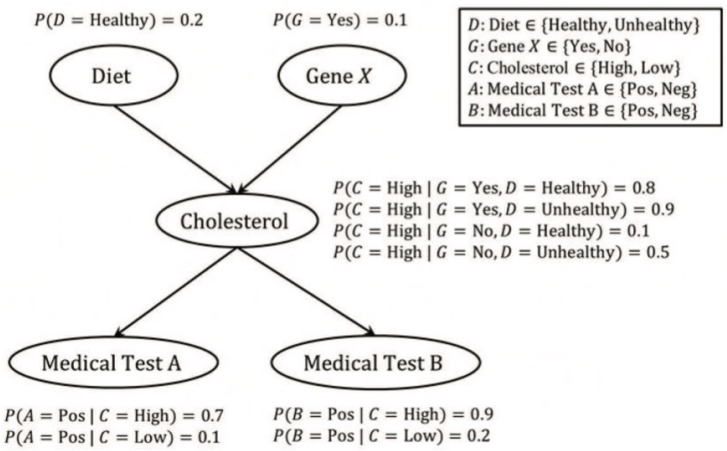


Figure 2: The Bayesian Network for Question 2.

Question 1

- If the person has high blood pressure, but exercises regularly and eats a healthy diet, to diagnose about heart disease (estimate the probabilities)

$$P(\text{HD=Yes}|\text{BP=High, D=Healthy, E=Yes})$$

v.s.

$$P(\text{HD=No}|\text{BP=High, D=Healthy, E=Yes})$$

$$P(\text{HD=Yes}|\text{BP=High, D=Healthy, E=Yes})$$

$$= \frac{P(\text{HD=Yes, BP=High, D=Healthy, E=Yes})}{P(\text{BP=High, D=Healthy, E=Yes})}$$

Denote by $\mathbf{U} = \{\text{BP, D, E}\}$

$$P(\text{HD}|\mathbf{U}) = \frac{P(\text{HD}, \mathbf{U})}{P(\mathbf{U})}$$

$$= \frac{P(\text{BP=High}|\text{HD=Yes}, \text{D=Healthy}, \text{E=Yes})P(\text{HD=Yes, D=Healthy, E=Yes})}{P(\text{BP=High, D=Healthy, E=Yes})}$$

Next slide

When HD is observed, BP is conditionally independent of E and D

$$P(\text{BP=High}|\text{HD=Yes}) = 0.85$$

$$P(\text{HD=Yes}|\text{BP=High, D=Healthy, E=Yes}) = \frac{P(\text{BP=High}|\text{HD=Yes})P(\text{HD=Yes}|\text{D=Healthy, E=Yes})}{P(\text{BP=High}|\text{D=Healthy, E=Yes})}$$

$$= \frac{0.85 \times 0.25}{P(\text{BP=High}|\text{D=Healthy, E=Yes})}$$

$$P(\text{HD=No}|\text{BP=High, D=Healthy, E=Yes}) = \frac{P(\text{BP=High}|\text{HD=No})P(\text{HD=No}|\text{D=Healthy, E=Yes})}{P(\text{BP=High}|\text{D=Healthy, E=Yes})}$$

$$= \frac{0.2 \times 0.75}{P(\text{BP=High}|\text{D=Healthy, E=Yes})}$$

$P(\text{HD=No}|\text{D=Healthy, E=Yes}) = 1 - P(\text{HD=Yes}|\text{D=Healthy, E=Yes}) = 0.75$

$$P(\text{HD=Yes}|\text{BP=High, D=Healthy, E=Yes}) = \frac{0.85 \times 0.25}{P(\text{BP=High}|\text{D=Healthy, E=Yes})}$$

$$P(\text{HD=No}|\text{BP=High, D=Healthy, E=Yes}) = \frac{0.2 \times 0.75}{P(\text{BP=High}|\text{D=Healthy, E=Yes})}$$

$$\frac{0.85 \times 0.25}{P(\text{BP=High}|\text{D=Healthy, E=Yes})} + \frac{0.2 \times 0.75}{P(\text{BP=High}|\text{D=Healthy, E=Yes})} = 1$$

$$\frac{0.85 \times 0.25 + 0.2 \times 0.75}{P(\text{BP=High}|\text{D=Healthy, E=Yes})} = 1$$

$$P(\text{BP=High}|\text{D=Healthy, E=Yes}) = 0.85 \times 0.25 + 0.2 \times 0.75$$

How to estimate $P(\text{BP=High}|\text{D=Healthy, E=Yes})$?

$$P(\text{HD=Yes}|\text{BP=High, D=Healthy, E=Yes}) = \frac{0.85 \times 0.25}{0.85 \times 0.25 + 0.2 \times 0.75} = 0.5862$$

$$P(\text{HD=No}|\text{BP=High, D=Healthy, E=Yes}) = \frac{0.2 \times 0.75}{0.85 \times 0.25 + 0.2 \times 0.75} = 0.4138$$

Question 2

- Consider the BBN below. Given that the outcomes of medical test A and medical test B for a specific patient are positive and negative, respectively. That is, A = Pos and B = Neg. Predict the probability that the patient has high cholesterol.

$$P(C = \text{High}|A = \text{Pos}, B = \text{Neg})$$

$$P(C = \text{High}|A = \text{Pos}, B = \text{Neg}) = \frac{P(A = \text{Pos}, B = \text{Neg}, C = \text{High})P(C = \text{High})}{P(A = \text{Pos}, B = \text{Neg})}$$

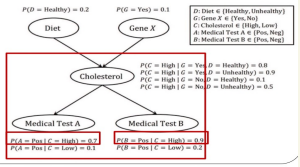
C is the parent node of both A and B

$$= \frac{P(A = \text{Pos}|C = \text{High})P(B = \text{Neg}|C = \text{High})P(C = \text{High})}{P(A = \text{Pos}, B = \text{Neg})}$$

$$= \frac{0.7 \times 0.1 \times P(C = \text{High})}{P(A = \text{Pos}, B = \text{Neg})}$$

$$P(B = \text{Pos}|C = \text{High}) = 0.9$$

$$P(B = \text{Neg}|C = \text{High}) = 0.1$$



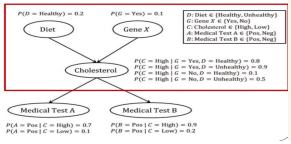
$$P(C = \text{High}) = \sum_{\alpha,\beta} P(C = \text{High}|D = \alpha, G = \beta)P(D = \alpha)P(G = \beta)$$

$$= P(C = \text{High}|D = \text{Healthy}, G = \text{Yes})P(D = \text{Healthy})P(G = \text{Yes}) + P(C = \text{High}|D = \text{Healthy}, G = \text{No})P(D = \text{Healthy})P(G = \text{No}) + P(C = \text{High}|D = \text{Unhealthy}, G = \text{Yes})P(D = \text{Unhealthy})P(G = \text{Yes}) + P(C = \text{High}|D = \text{Unhealthy}, G = \text{No})P(D = \text{Unhealthy})P(G = \text{No})$$

$$= 0.8 \times 0.2 \times 0.1 + 0.1 \times 0.2 \times 0.9 + 0.9 \times 0.8 \times 0.1 + 0.5 \times 0.8 \times 0.9$$

$$= 0.47$$

Thus, $P(C = \text{High}|A = \text{Pos}, B = \text{Neg}) = \frac{0.7 \times 0.1 \times P(C = \text{High})}{P(A = \text{Pos}, B = \text{Neg})} = \frac{0.7 \times 0.1 \times 0.47}{P(A = \text{Pos}, B = \text{Neg})} = \frac{0.0329}{P(A = \text{Pos}, B = \text{Neg})}$



$$P(C = \text{Low}|A = \text{Pos}, B = \text{Neg}) = \frac{P(A = \text{Pos}, B = \text{Neg}|C = \text{Low})P(C = \text{Low})}{P(A = \text{Pos}, B = \text{Neg})}$$

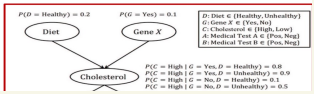
$$P(C = \text{High}) = 0.47$$

$$P(C = \text{Low}) = 1 - 0.47 = 0.53$$

$$= \frac{P(A = \text{Pos}|C = \text{Low})P(B = \text{Neg}|C = \text{Low})P(C = \text{Low})}{P(A = \text{Pos}, B = \text{Neg})}$$

$$= \frac{0.1 \times 0.8 \times 0.53}{P(A = \text{Pos}, B = \text{Neg})}$$

$$= \frac{0.0424}{P(A = \text{Pos}, B = \text{Neg})}$$



$$P(C = \text{High}) = \sum_{\alpha,\beta} P(D = \alpha, G = \beta, C = \text{High}) = \sum_{\alpha,\beta} P(C = \text{High}|D = \alpha, G = \beta)P(D = \alpha, G = \beta) = \sum_{\alpha,\beta} P(C = \text{High}|D = \alpha, G = \beta)P(D = \alpha)P(G = \beta)$$

$\alpha \in \{\text{Healthy, Unhealthy}\}$

$\beta \in \{\text{Yes, No}\}$

$$P(C = \text{Low}|A = \text{Pos}, B = \text{Neg}) + P(C = \text{High}|A = \text{Pos}, B = \text{Neg}) = 1 \quad \frac{0.0424 + 0.0329}{P(A = \text{Pos}, B = \text{Neg})} = 1 \quad \Rightarrow \quad P(A = \text{Pos}, B = \text{Neg}) = 0.0753$$

$$\frac{0.0424}{P(A = \text{Pos}, B = \text{Neg})} \qquad \frac{0.0329}{P(A = \text{Pos}, B = \text{Neg})}$$

Thus, $P(C = \text{High}|A = \text{Pos}, B = \text{Neg}) = \frac{0.0329}{0.0753} = 0.44$

SC4000/CZ4041/CE4041: Machine Learning
Lecture 5 Tutorial Question Sets

Question 1: Consider the data set shown in Table 1 for a binary classification problem.

Table 1: Data set for Question 1.

A	B	Class Label
M	F	+
F	T	+
T	T	+
M	F	-
M	F	-
F	F	-
N	F	-
N	T	-
T	T	-
T	F	-

1. Calculate the information gain when splitting on A and B (using multi-way split on A). Which feature would the decision tree induction algorithm choose?

2. Calculate the gain ratio when splitting on A and B (using multi-way split on A). Which feature would the decision tree induction algorithm choose?

Entropy & Information Gain

$$\text{Entropy}(t) = -\sum_{c} P(y=c;t)\log_2 P(y=c;t)$$

- Suppose a parent node t is split into P partitions (children)
- Information Gain:

$$\Delta_{\text{info}} = \text{Entropy(t)} - \sum_{j=1}^{P} \frac{n_j}{n}\text{Entropy}(j)$$

(n_j: Number of examples at child j; n: Number of examples at node t)

- To choose a feature whose test condition maximizes the gain

Question 1.1

Table 1: Data set for Question 1.

A	B	Class Label
M	F	+
F	T	+
T	T	+
M	F	-
M	F	-
F	F	-
N	F	-
N	T	-
T	T	-
T	F	-

Task: calculate information gain when splitting on A (multi-way) and B. Which feature to choose?

	Parent
+	3
−	7

	$A=T$	$A=F$	$A=M$	$A=N$
+	1	1	1	0
−	2	1	2	2

Split on A

	$B=T$	$B=F$
+	2	1
−	2	5

Split on B

$$\text{Entropy(Parent)} = -\left(\frac{3}{10}\right)\log_2\left(\frac{3}{10}\right) - \left(\frac{7}{10}\right)\log_2\left(\frac{7}{10}\right) = 0.8813$$

$$\text{Entropy}(A=T) = -\left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) = 0.9183$$

$$\text{Entropy}(A=F) = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = 1$$

$$\text{Entropy}(A=M) = -\left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) = 0.9183$$

$$\text{Entropy}(A=N) = -\left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right) - \left(\frac{2}{2}\right)\log_2\left(\frac{2}{2}\right) = 0$$

$$\text{Entropy}(\text{Split}_A) = \left(\frac{3}{10}\right)\times 0.9183 + \left(\frac{2}{10}\right)\times 1 + \left(\frac{3}{10}\right)\times 0.9183 + \left(\frac{2}{10}\right)\times 0 = 0.7510$$

$$\Delta_{\text{info}}(A) = 0.8813 - 0.7510 = 0.1303$$

Split on A

$$\text{Entropy(Parent)} = -\left(\frac{3}{10}\right)\log_2\left(\frac{3}{10}\right) - \left(\frac{7}{10}\right)\log_2\left(\frac{7}{10}\right) = 0.8813$$

	Parent
+	3
−	7

$$\text{Entropy}(B=T) = -\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) = 1$$

$$\text{Entropy}(B=F) = -\left(\frac{1}{6}\right)\log_2\left(\frac{1}{6}\right) - \left(\frac{5}{6}\right)\log_2\left(\frac{5}{6}\right) = 0.65$$

	$B=T$	$B=F$
+	2	1
−	2	5

Split on B

$$\text{Entropy}(\text{Split}_B) = \left(\frac{4}{10}\right)\times 1 + \left(\frac{6}{10}\right)\times 0.65 = 0.79$$

$$\Delta_{\text{info}}(B) = 0.8813 - 0.79 = 0.0913 \quad < \quad \Delta_{\text{info}}(A) = 0.1303$$ ✓

Question 1.2: Calculate Gain Ratio

- Suppose a parent node t is split into P partitions (children)

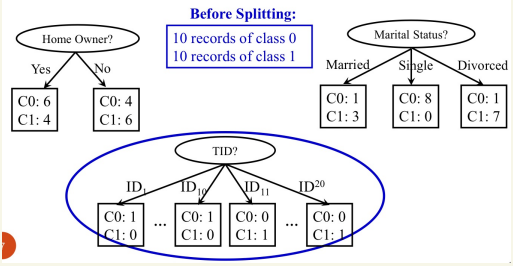
$$\text{Gain Ratio} = \frac{\Delta_{\text{info}}}{\text{SplitINFO}}$$

$$\text{where SplitINFO} = -\sum_{j=1}^{P}\frac{n_j}{n}\log_2\frac{n_j}{n}$$

(n_j: The number of records in partition j)

Why Gain Ratio?

- Disadvantage: tends to prefer splits that result in large number of partitions, each being small but pure



Question 1.2

$$\Delta_{\text{info}}(A) = 0.1303 \qquad \Delta_{\text{info}}(B) = 0.0913$$

$$\text{Gain Ratio} = \frac{\Delta_{\text{info}}}{\text{SplitINFO}} \quad \text{where SplitINFO} = -\sum_{i=1}^{P}\frac{n_i}{n}\log_2\frac{n_i}{n}$$

$A=T$	$A=F$	$A=M$	$A=N$
3	2	3	2

$$\text{SplitINFO}(A) = -\left(\frac{3}{10}\right)\log_2\left(\frac{3}{10}\right) - \left(\frac{2}{10}\right)\log_2\left(\frac{2}{10}\right) - \left(\frac{3}{10}\right)\log_2\left(\frac{3}{10}\right) - \left(\frac{2}{10}\right)\log_2\left(\frac{2}{10}\right) = 1.9710$$

$$\text{GainRatio}_A = \frac{\Delta_{\text{info}}(A)}{\text{SplitINFO}(A)} = \frac{0.1303}{1.9710} = 0.0661$$

$B=T$	$B=F$
4	6

$$\text{SplitINFO}(B) = -\left(\frac{4}{10}\right)\log_2\left(\frac{4}{10}\right) - \left(\frac{6}{10}\right)\log_2\left(\frac{6}{10}\right) = 0.9710$$

$$\text{GainRatio}_B = \frac{\Delta_{\text{info}}(B)}{\text{SplitINFO}(B)} = \frac{0.0913}{0.9710} = 0.094$$ ✓

Q1) $e'(t) = e(t) + N \times k = (3+2) + 2 \times 0.5 = 6$

case1 $e'(t) = 7 + 1 \times 0.5 = 7.5$, not prune

left

case2 $e'(t) = (3+2) + 2 \times 0.5 = 6$, prune

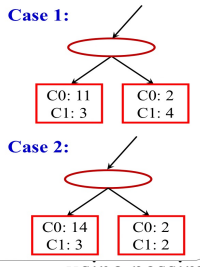
$e'(t) = 5 + 1 \times 0.5 = 5.5$

Question 1

- Pessimistic error?

$e'(T) = e(T) + N \times 0.5$

PRUNE?



00/CZ4041/CE4041: Machine Learning
Lecture 6 Tutorial Question Sets

Given the two cases shown in Figure 1, should we perform post-pruning using pessimistic error for each of them?

C0: 13
C1: 7

C0: 16
C1: 5

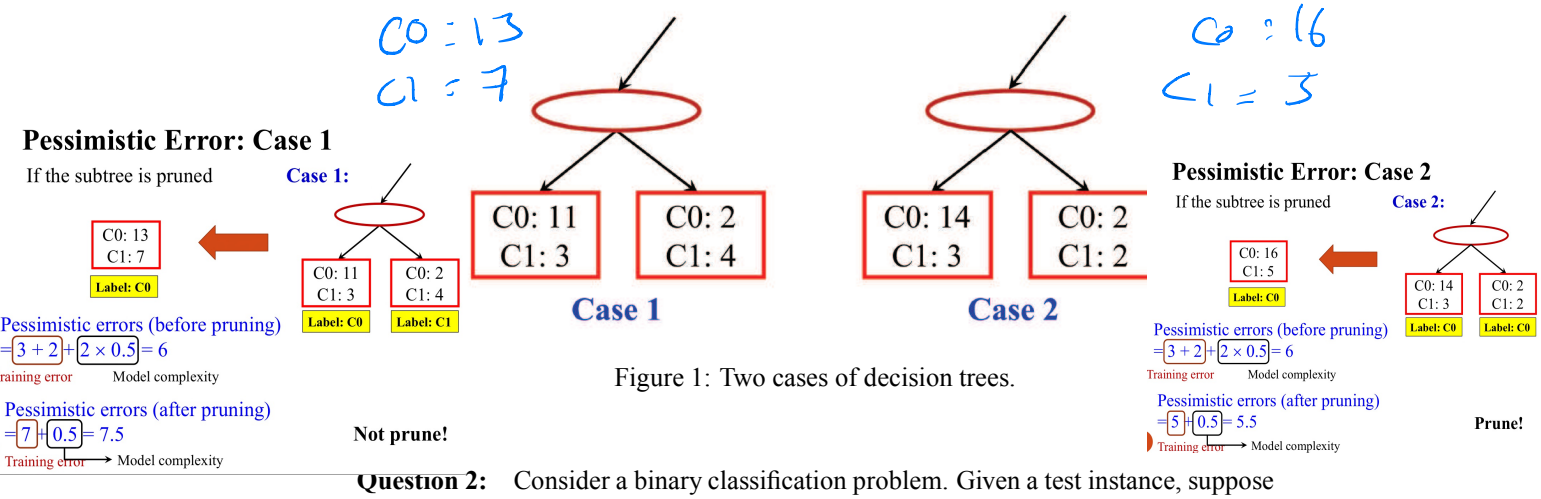


Figure 1: Two cases of decision trees.

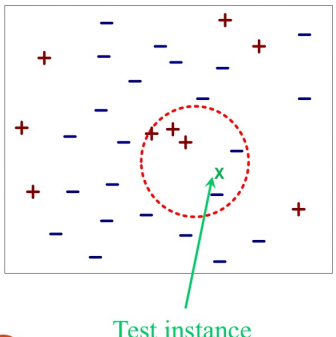
Question 2: Consider a binary classification problem. Given a test instance, suppose the 5 nearest neighbors of the test instance in the training data set are shown in Table 1. Use a 5-NN classifier with the distance-weighted voting approach to make a prediction on the test instance.

Table 1: The 5 nearest (training) neighbors to a test instance.

Training instance index	Class label	Distance to the test instance
1	+	3
2	+	3.5
3	+	4
4	−	1.5
5	−	2

Question 2

Consider a binary classification problem, and a 5-NN classifier



Retrieved instances →

Training instances	Class label	Distance to test instance
1	+	3
2	+	3.5
3	+	4
4	-	1.5
5	-	2

- Majority voting:

$+: 3 > -: 2$

- Distance-weighted voting:

Question 2 (cont.)

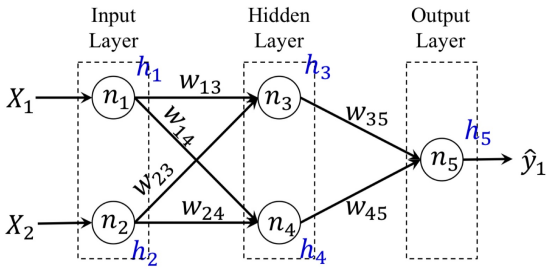
Training record	Class label	Distance to test record
1	+	3
2	+	3.5
3	+	4
4	-	1.5
5	-	2

Distance−Weighted voting for +:

$$\left(\frac{1}{3}\right)^2 + \left(\frac{1}{3.5}\right)^2 + \left(\frac{1}{4}\right)^2 = 0.2552$$

Distance−Weighted voting for −:

$$\left(\frac{1}{2}\right)^2 + \left(\frac{1}{1.5}\right)^2 = 0.6944$$



SC4000 Machine Learning Tutorial
Artificial Neural Networks

Question 1: On Slide 64 of Lecture 7, we have shown how to use backpropagation to update the parameters of ANN with one initialization setting for $\mathbf{w}$. Suppose now we initialize $\mathbf{w}$ with another set of values: $w_{13} = -1$, $w_{14} = -1$, $w_{23} = -1$, $w_{24} = -1$, $w_{35} = -1$, and $w_{45} = -1$. Run one epoch (i.e., run through the whole training dataset once), to show how the parameters are updated at each iteration.

Question 2: Consider a 2-dimensional dataset for three-class classification by ANN, as shown in Figure 1. Which ANN model as shown in Figure 2 is proper to solve the classification problem? Why?

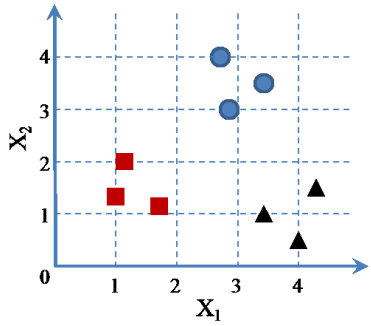
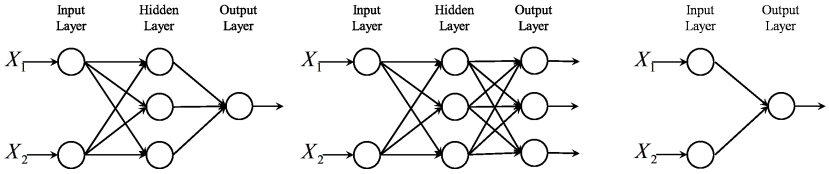


Figure 1: Dataset for Question 2.



(a) ANN model A. (b) ANN model B. (c) ANN model C.

Figure 2: Different ANN structures for Question 2.

Question 3: Compute the derivative of the sigmoid function $f(z) = \frac{1}{1+e^{-z}}$ w.r.t. z.

Question 1

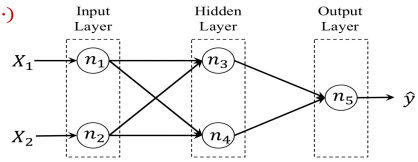
- In Lecture 7, we showed how to use backpropagation to update the parameters of ANN with one initialization setting for w.
- Suppose now we initialize w with another set of values: $w_{13} = -1$, $w_{14} = -1$, $w_{23} = -1$, $w_{24} = -1$, $w_{35} = -1$, and $w_{45} = -1$.
- Run one epoch (i.e., run through the whole training dataset once), to show how the parameters are updated at each iteration.

Question 1

$\lambda = 0.4$, $\theta = 0$

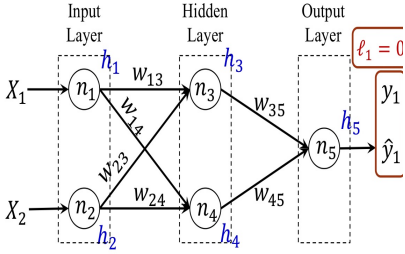
Sign function as $a(\cdot)$

X_1	X_2	y
0	0	-1
1	0	1
0	1	1
1	1	1



- **Initialization 1 (lecture notes):**
  $(w_{13} = 1, w_{14} = 1, w_{23} = 1, w_{24} = 1, w_{35} = 1, w_{45} = 1)$
- **Initialization 2:**
  $(w_{13} = -1, w_{14} = -1, w_{23} = -1, w_{24} = -1, w_{35} = -1, w_{45} = -1)$

X_1	X_2	y
0	0	-1
1	0	1
0	1	1
1	1	1



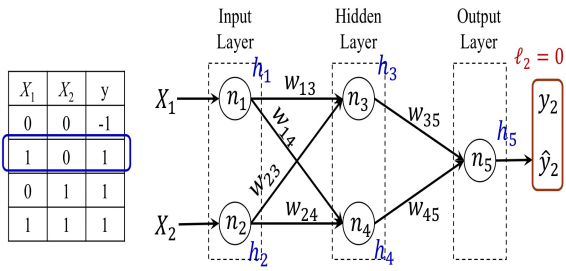
Forward pass:

For the 1st example: $h_1 = 0$ and $h_2 = 0$

$h_3 = \text{sign}(0 \times (-1) + 0 \times (-1)) = 1$ and $h_4 = \text{sign}(0 \times (-1) + 0 \times (-1)) = 1$

Then $\hat{y}_1 = h_5 = \text{sign}(1 \times (-1) + 1 \times (-1)) = -1$

X_1	X_2	y
0	0	-1
1	0	1
0	1	1
1	1	1

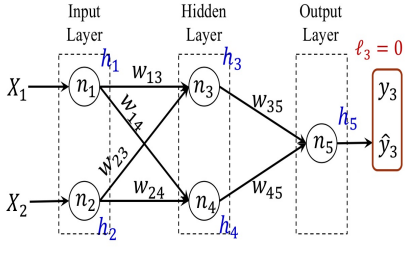


For the 2nd example: $h_1 = 1$ and $h_2 = 0$

$h_3 = \text{sign}(1 \times (-1) + 0 \times (-1)) = -1$ and $h_4 = \text{sign}(1 \times (-1) + 0 \times (-1)) = -1$

Then $\hat{y}_2 = h_5 = \text{sign}((-1) \times (-1) + (-1) \times (-1)) = 1$

X_1	X_2	y
0	0	-1
1	0	1
0	1	1
1	1	1

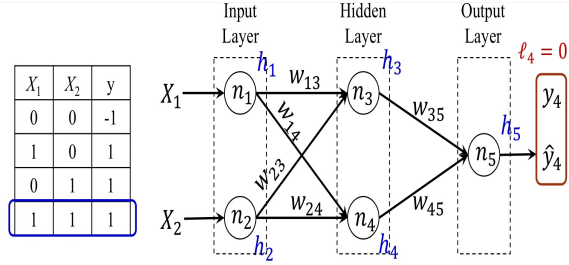


For the 3rd example: $h_1 = 0$ and $h_2 = 1$

$h_3 = \text{sign}(0 \times (-1) + 1 \times (-1)) = -1$ and $h_4 = \text{sign}(0 \times (-1) + 1 \times (-1)) = -1$

Then $\hat{y}_3 = h_5 = \text{sign}((-1) \times (-1) + (-1) \times (-1)) = 1$

X_1	X_2	y
0	0	-1
1	0	1
0	1	1
1	1	1



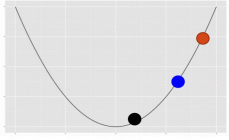
For the 4th example: $h_1 = 1$ and $h_2 = 1$

$h_3 = \text{sign}(1 \times (-1) + 1 \times (-1)) = -1$ and $h_4 = \text{sign}(1 \times (-1) + 1 \times (-1)) = -1$

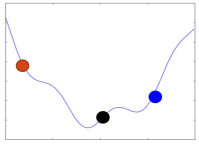
Then $\hat{y}_4 = h_5 = \text{sign}((-1) \times (-1) + (-1) \times (-1)) = 1$

Question 1: Conclusion

- Different initializations may lead to different convergence iterations
- Different initializations may lead to different local optima



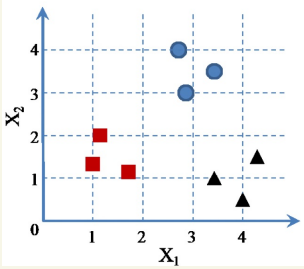
Convex

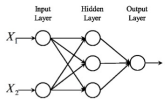


Non-convex

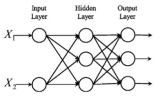
Question 2

- Consider a 2-dimensional dataset for three-class classification by ANN, as shown in Figure 1. Which ANN model as shown in Figure 2 is proper to solve the classification problem? Why?

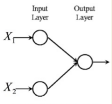




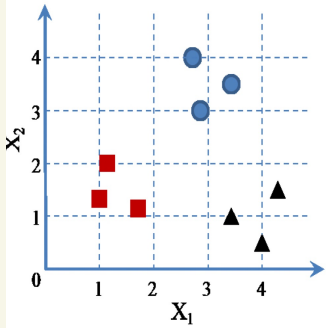
(a) ANN model *A*.



(b) ANN model *B*.

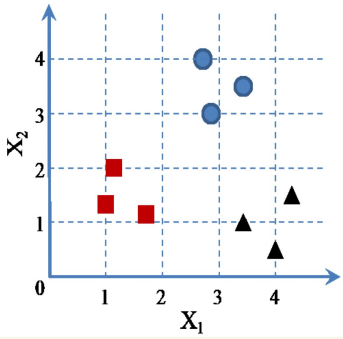


(c) ANN model *C*.



The instances of different classes are not able to be linearly separated

Therefore, the Perceptron model does not work



There are three different classes

Design Issues for ANN

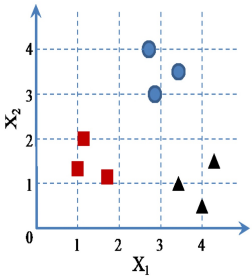
- The number of nodes in the input layer
  - Assign an input node to each numerical or binary input variable
- The number of nodes in the output layer
  - Binary class problem → single node
  - C-class problem → C output nodes
- We output a one-hot encoding of the class
  - [1, 0, 0] for class 1
  - [0, 1, 0] for class 2

[Optional] Multi-class activation

- In a multi-class classification problem with C classes, we should output a one-hot encoding of the class
  - [1, 0, 0] for class 1
  - [0, 1, 0] for class 2
- We take the pre-activation outputs, $z_{L,1}, z_{L,2}$, and $z_{L,3}$, and feed them to a softmax function.

$$\hat{y}_i = \frac{\exp(z_{L,i})}{\sum_j \exp(z_{L,j})}$$

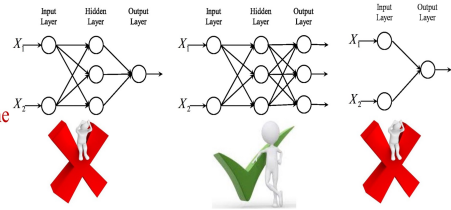
- This results in $1 > \hat{y}_i > 0$ and $\sum_i \hat{y}_i = 1$



There are three different classes

Therefore, there should be three output nodes in the output layer

Question 2 (cont.)



Question 3

- Compute the derivative of the sigmoid function w.r.t. z:

$$f(z) = \frac{1}{1+e^{-z}}$$

- Denote $y = 1 + e^{-z}$, i.e., y is a function of z
- And then f is a function of y: $f = \frac{1}{y}$
- Based on the chain rule:

$$\frac{\partial f(z)}{\partial z} = \frac{\partial f}{\partial y}\frac{\partial y}{\partial z} = \left(\frac{-1}{y^2}\right)\left(\cancel{\frac{\partial 1}{\partial z}}^{0} + \boxed{\frac{\partial e^{-z}}{\partial z}}^{-e^{-z}}\right) = \frac{e^{-z}}{y^2}$$

$$\frac{\partial f(z)}{\partial z} = \frac{\partial f}{\partial y}\frac{\partial y}{\partial z} = \frac{e^{-z}}{y^2}$$

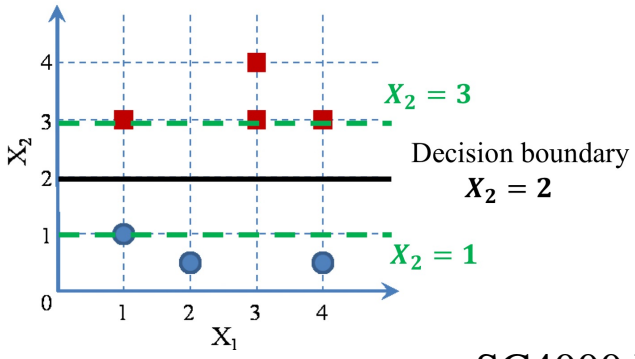
$$y = 1 + e^{-z}$$

$$\frac{\partial f(z)}{\partial z} = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1}{1+e^{-z}}\left(\frac{e^{-z}}{1+e^{-z}}\right)$$

$$\boxed{f(z) = \frac{1}{1+e^{-z}}} \quad = f(z)\left(1 - \frac{1}{1+e^{-z}}\right)$$

$$= f(z)(1 - f(z))$$

Question 1



SC4000 Machine Learning Tutorial
Support Vector Machines and Linear Regression

Question 1: Consider a 2-dimensional dataset for two-class classification by SVM, as shown in Figure 1, where the red "square" and blue "circle" denote the positive and negative classes respectively. Is this dataset separable by a linear SVM classifier? If no, why? If yes, what is the decision boundary of the linear SVM? And what are the pair of parallel hyperplanes associated with the decision boundary? (No need to provide proofs)

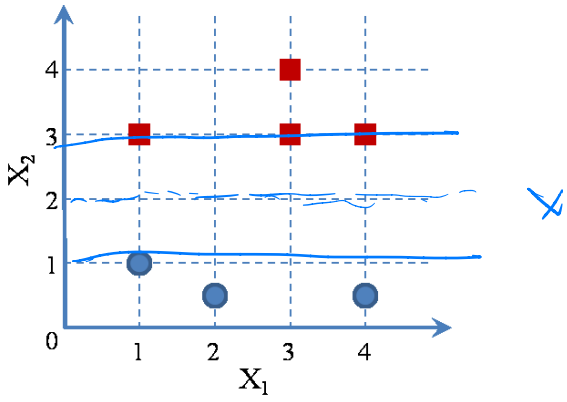


Figure 1: Dataset for Question 1.

Question 2: Please induce why the two parallel hyperplanes of a decision boundary,

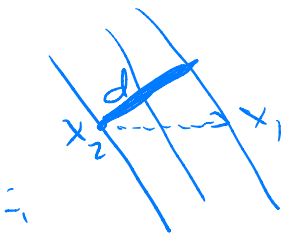
$$\mathbf{w} \cdot \mathbf{x} + b = 0,$$

can be written as

$$\mathbf{w} \cdot \mathbf{x} + b = \bar{k}, \text{ and } \mathbf{w} \cdot \mathbf{x} + b = -\bar{k},$$

where $\bar{k} > 0$, respectively.

Question 3: In an SVM, we have two support vectors $(2, 3)$ from class 1 and $(-1, 4)$ from class 2. The separating hyperplane can be written as $3x_1 + x_2 + b = 0$. Compute the margin of separation.



Question 2

The two parallel hyperplanes passing the closest circle(s) and square(s) can be written as

$\boldsymbol{w} \cdot \boldsymbol{x} + b = k$, where $k > 0$

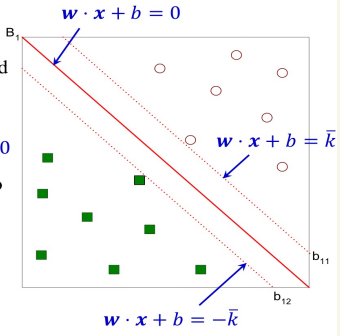
$\boldsymbol{w} \cdot \boldsymbol{x} + b = k'$, where $k' < 0$

It can be shown that, these two parallel hyperplanes can be further rewritten as

$\boldsymbol{w} \cdot \boldsymbol{x} + b = \bar{k}$

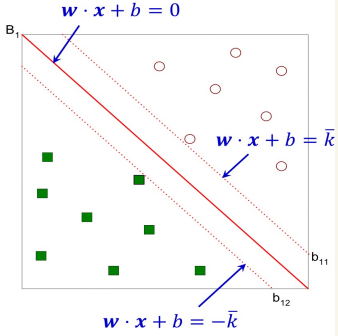
$\boldsymbol{w} \cdot \boldsymbol{x} + b = -\bar{k}$

where $\bar{k} > 0$



Question 2

- $\boldsymbol{w}$ determines the orientation (slope) of the decision boundary.
- The support vectors determine how the decision boundary moves in parallel motion.
- Together, they determine b.

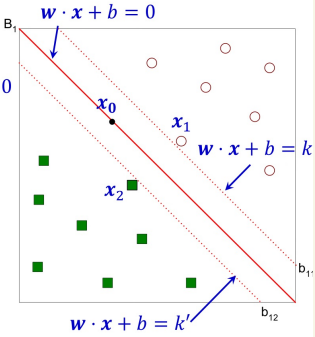


Question 2 (cont.)

b_{11}: $\boldsymbol{w} \cdot \boldsymbol{x}_1 + b = k$, where $k > 0$

b_{12}: $\boldsymbol{w} \cdot \boldsymbol{x}_2 + b = k'$, where $k' < 0$

Given two support vectors (or two points on b_{11} and b_{12} repsectively), I can choose b such that $k = -k'$

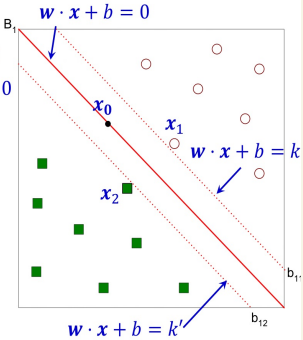


Question 2 (cont.)

b_{11}: $\boldsymbol{w} \cdot \boldsymbol{x}_1 + b = k$, where $k > 0$

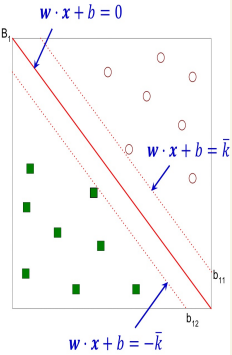
b_{12}: $\boldsymbol{w} \cdot \boldsymbol{x}_2 + b = k'$, where $k' < 0$

$$\boldsymbol{w} \cdot \boldsymbol{x}_1 + b = -(\boldsymbol{w} \cdot \boldsymbol{x}_2 + b)$$
$$2b = -\boldsymbol{w} \cdot \boldsymbol{x}_1 - \boldsymbol{w} \cdot \boldsymbol{x}_2$$
$$b = -\frac{1}{2}\boldsymbol{w} \cdot (\boldsymbol{x}_1 + \boldsymbol{x}_2)$$



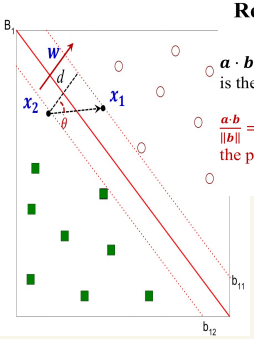
Question 3

- Two support vectors: (2, 3) and (-1, 4) from the two classes, respectively.
- Decision boundary: $\boldsymbol{w} \cdot \boldsymbol{x} + b = 3x_1 + x_2 + b = 0$
- What is the margin?



Question 3

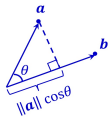
- From the lecture, $\boldsymbol{w}$ is orthogonal to the decision boundary.
- All we need is to find the length of projection of the vector $(\boldsymbol{x}_1 - \boldsymbol{x}_2)$ onto the direction of $\boldsymbol{w}$.



Review: Geometry of Inner Products

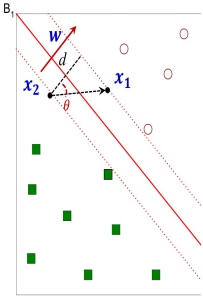
$\boldsymbol{a} \cdot \boldsymbol{b} = \|\boldsymbol{a}\|\|\boldsymbol{b}\| \cos\theta$, where θ is the angle between $\boldsymbol{a}$ and $\boldsymbol{b}$

$\frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{b}\|} = \|\boldsymbol{a}\| \cos\theta$ is the length of the projection of $\boldsymbol{a}$ onto $\boldsymbol{b}$



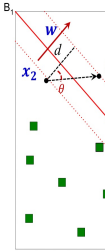
Question 3

- We know that
  $\boldsymbol{w} \cdot (\boldsymbol{x}_1 - \boldsymbol{x}_2) = \|\boldsymbol{w}\| \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \cos(\theta)$
  $= \|\boldsymbol{w}\| d$ or $-\|\boldsymbol{w}\| d$
- Thus, $d = \left|\frac{\boldsymbol{w} \cdot (\boldsymbol{x}_1 - \boldsymbol{x}_2)}{\|\boldsymbol{w}\|}\right|$
- $\boldsymbol{x}_1 - \boldsymbol{x}_2 = (2,3) - (-1,4)$
  $= (3,-1)$



Question 3

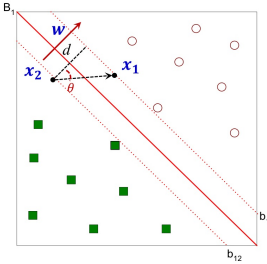
- Thus, $d = \left|\frac{\boldsymbol{w} \cdot (\boldsymbol{x}_1 - \boldsymbol{x}_2)}{\|\boldsymbol{w}\|}\right|$
- $\boldsymbol{x}_1 - \boldsymbol{x}_2 = (2,3) - (-1,4)$
  $= (3,-1)$
- We know $\boldsymbol{w} = (3,1)$
- $|\boldsymbol{w} \cdot (\boldsymbol{x}_1 - \boldsymbol{x}_2)| = |3 \times 3 - 1 \times 1| = 8$
- $\|\boldsymbol{w}\| = \sqrt{10}$
- $d = \frac{8}{\sqrt{10}}$



Question 3

- Why can't we directly use the equation $\|\boldsymbol{w}\| d = 2$?
- That equation is only valid when $\boldsymbol{w}$ is properly rescaled.
- That is, when
  $\boldsymbol{w} \cdot \boldsymbol{x}_1 + b = 1$
  $\boldsymbol{w} \cdot \boldsymbol{x}_2 + b = -1$

Can only use if k and $-k$



Q4: Regularized Linear Regression

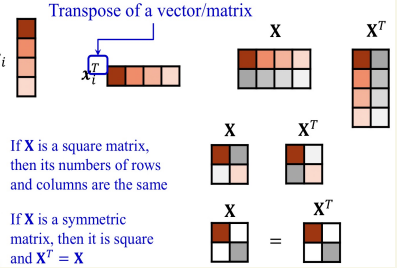
- To solve the unconstrained minimization problem, we can set the derivative of $\mathcal{L}(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$ w.r.t. $\boldsymbol{w}$ to **zero**

$$\frac{\partial\left(\mathcal{L}(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2\right)}{\partial \boldsymbol{w}} = \frac{\partial\left(\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}\cdot\boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2}\boldsymbol{w}\cdot\boldsymbol{w}\right)}{\partial \boldsymbol{w}} = \mathbf{0}$$

- We can obtain a closed-form solution

Some Concepts: Review

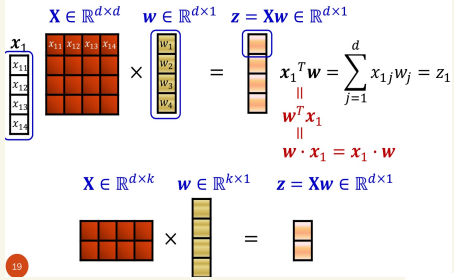
Transpose of a vector/matrix



Some Concepts: Review

- The transpose of $\mathbf{XY}$: $(\mathbf{XY})^T = \mathbf{Y}^T\mathbf{X}^T$
- The transpose of $\mathbf{Xw}$: $(\mathbf{Xw})^T = \boldsymbol{w}^T\mathbf{X}^T$
- The transpose of $\boldsymbol{x}^T\boldsymbol{w}$: $(\boldsymbol{x}^T\boldsymbol{w})^T = \boldsymbol{w}^T\boldsymbol{x}$
- The transpose of a scalar is the scalar itself: $x^T = x$

Some Concepts: Review (cont.)

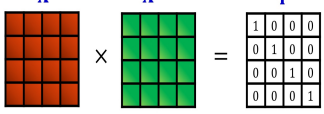


19

Some Concepts: Review (cont.)

- For a square matrix $\mathbf{X}$, if $\mathbf{X}$ is invertible, then

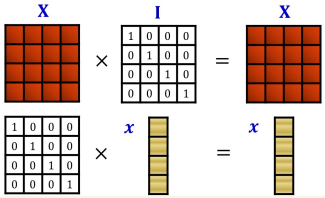
$\mathbf{X}\mathbf{X}^{-1} = \mathbf{I}$ ← Identity matrix



Some Concepts: Review (cont.)

- Any vector (or matrix) $\boldsymbol{x}$ (or $\mathbf{X}$) times identity matrix $\mathbf{I}$ equals to the vector (or matrix) itself

$\mathbf{I}\boldsymbol{x} = \boldsymbol{x}\ (\boldsymbol{x}^T\mathbf{I} = \boldsymbol{x}^T)$ OR $\mathbf{X}\mathbf{I} = \mathbf{X}\ (\mathbf{I}\mathbf{X} = \mathbf{X})$



Question 4

- Denote by $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)^T$

$$\mathbf{X} = \begin{pmatrix} x_{10} & \cdots & x_{N0} \\ \vdots & \ddots & \vdots \\ x_{1d} & \cdots & x_{Nd} \end{pmatrix}^T = \begin{pmatrix} x_{10} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{N0} & \cdots & x_{Nd} \end{pmatrix}$$

($\boldsymbol{x}_1$ marks the first column of $\mathbf{X}$ and the first row of $\mathbf{X}^T$)

- And by $\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$

How to get this closed-form solution?

- The closed-form solution for $\boldsymbol{w}$:

$$\boldsymbol{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{y}$$

23

$$\frac{\partial\left(\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}\cdot\boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2}\boldsymbol{w}\cdot\boldsymbol{w}\right)}{\partial \boldsymbol{w}} = \mathbf{0}$$

$$\frac{\partial \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}\cdot\boldsymbol{x}_i - y_i)^2}{\partial \boldsymbol{w}} = \frac{1}{2}\sum_{i=1}^{N} 2(\boldsymbol{w}\cdot\boldsymbol{x}_i - y_i)\frac{\partial(\boldsymbol{w}\cdot\boldsymbol{x}_i - y_i)}{\partial \boldsymbol{w}} = \sum_{i=1}^{N}(\boldsymbol{w}\cdot\boldsymbol{x}_i - y_i)\boldsymbol{x}_i$$

$$\frac{\partial\left(\frac{\lambda}{2}\boldsymbol{w}\cdot\boldsymbol{w}\right)}{\partial \boldsymbol{w}} = \frac{\lambda}{2}\times 2\boldsymbol{w} = \lambda\boldsymbol{w}$$

$$\sum_{i=1}^{N}(\boldsymbol{w}\cdot\boldsymbol{x}_i - y_i)\boldsymbol{x}_i + \lambda\boldsymbol{w} = \mathbf{0}$$

$$\sum_{i=1}^{N}(\boldsymbol{w}\cdot\boldsymbol{x}_i)\boldsymbol{x}_i - \sum_{i=1}^{N} y_i\boldsymbol{x}_i + \lambda\boldsymbol{w} = \mathbf{0}$$

$$\sum_{i=1}^{N}(\boldsymbol{w}\cdot\boldsymbol{x}_i)\boldsymbol{x}_i - \sum_{i=1}^{N} y_i\boldsymbol{x}_i + \lambda\boldsymbol{w} = \mathbf{0}$$

$$\boldsymbol{x}_i(\boldsymbol{w}\cdot\boldsymbol{x}_i) = \boldsymbol{x}_i(\boldsymbol{x}_i^T\boldsymbol{w}) = (\boldsymbol{x}_i\boldsymbol{x}_i^T)\boldsymbol{w}$$

$$\left(\sum_{i=1}^{N}(\boldsymbol{x}_i\boldsymbol{x}_i^T)\right)\boldsymbol{w} - \sum_{i=1}^{N} y_i\boldsymbol{x}_i + \lambda\mathbf{I}\boldsymbol{w} = \mathbf{0}$$

Identity matrix

A matrix of $(d+1)$ by $(d+1)$ size, where $\boldsymbol{x}_i$ is a column vector of $(d+1)$ dimensions

$$\boldsymbol{x}_i \quad \boldsymbol{x}_i^T$$

$$\begin{bmatrix} x_{i0} \\ x_{i1} \\ \cdots \\ x_{id} \end{bmatrix} \quad [x_{i0} \quad x_{i1} \quad \ldots \quad x_{id}]$$

$$\boldsymbol{x}_i\boldsymbol{x}_i^T = \begin{pmatrix} x_{i0}\times x_{i0} & \cdots & x_{i0}\times x_{id} \\ \vdots & \ddots & \vdots \\ x_{id}\times x_{i0} & \cdots & x_{id}\times x_{id} \end{pmatrix}$$

$$\sum_{i=1}^{N}(\boldsymbol{x}_i\boldsymbol{x}_i^T) = \begin{pmatrix} \sum_{i=1}^{N} x_{i0}\times x_{i0} & \cdots & \sum_{i=1}^{N} x_{i0}\times x_{id} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{N} x_{id}\times x_{i0} & \cdots & \sum_{i=1}^{N} x_{id}\times x_{id} \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} x_{10} & \cdots & x_{N0} \\ \vdots & \ddots & \vdots \\ x_{1d} & \cdots & x_{Nd} \end{pmatrix}^T = \begin{pmatrix} x_{10} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{N0} & \cdots & x_{Nd} \end{pmatrix}$$

($\boldsymbol{x}_1$ marks the first column of $\mathbf{X}$ and the first row of $\mathbf{X}^T$)

$$\sum_{i=1}^{N}(\boldsymbol{x}_i\boldsymbol{x}_i^T) = \mathbf{X}^T\mathbf{X}$$

Therefore $\left(\sum_{i=1}^{N}(\boldsymbol{x}_i\boldsymbol{x}_i^T)\right)\boldsymbol{w} - \sum_{i=1}^{N} y_i\boldsymbol{x}_i + \lambda\mathbf{I}\boldsymbol{w} = \mathbf{0}$

$$(\mathbf{X}^T\mathbf{X})\boldsymbol{w} - \mathbf{X}^T\boldsymbol{y} + \lambda\mathbf{I}\boldsymbol{w} = \mathbf{0}$$

Always invertible as long as λ is positive

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\boldsymbol{w} = \mathbf{X}^T\boldsymbol{y}$$

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\boldsymbol{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{y}$$

$$\mathbf{I}\boldsymbol{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{y}$$

$$\boldsymbol{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{y}$$

Question 1

- Suppose there are 3 base classifiers, each classifier has error rate, $\varepsilon = 0.65$ or accuracy $acc = 0.35$
- Consider to combine the 3 base classifiers to make a prediction on a test instance using a majority vote
- Therefore, probability that the ensemble classifier makes a wrong prediction is:

$$\sum_{i=2}^{3} \binom{3}{i} \varepsilon^i (1-\varepsilon)^{3-i} = 3 \times 0.65^2 \times 0.35 + 1 \times 0.65^3 \times 1 = 0.71825$$

SC4000 Machine Learning Tutorial
Ensemble Learning

Question 1: Why is the condition "the base classifiers should do better than a classifier that performs random guessing" necessary for ensemble learning?

Question 2: Suppose we have trained 5 base binary classifiers: f_1, f_2, f_3, f_4 and f_5. Their predictions on a validation dataset are shown in Table 1, where the last column denotes the ground-truth class labels. Which base classifiers would you choose to construct an ensemble learner?

The Binomial Distribution

- This is the general expression for the binomial distribution.
- We perform N experiments. In each experiment, the probability of observing a negative outcome is ϵ. What is the probability for observing exactly M negative outcomes?

$$P(N, M) = \binom{N}{M} \varepsilon^M (1-\varepsilon)^{N-M}$$

Table 1: Data set for Question 2.

ID	f_1	f_2	f_3	f_4	f_5	Ground Truth
P1	+	+	-	-	+	+
P2	+	+	-	+	-	+
P3	-	-	+	+	-	+
P4	-	-	+	-	+	+
P5	-	-	+	+	-	-
P6	-	-	-	+	+	+
P7	+	+	+	+	-	+
P8	-	+	+	-	+	-
P9	+	+	-	+	+	+
P10	-	-	-	+	-	-

Applying That Formula …

- Suppose there are N (odd) independent base classifiers, each of which has the same error rate ε
  - Therefore, probability that the ensemble classifier makes a wrong prediction is:

$$P(N) = \sum_{i=\frac{N+1}{2}}^{N} \binom{N}{i} \varepsilon^i (1-\varepsilon)^{N-i}$$

There are i classifiers that make wrong predictions

The rest $N-i$ classifiers making correct prediction

At least more than half of the base classifiers

All possible combinations

Another Application

- Alternatively, let p be the probability that a single classifier makes the correct decision. The probability that the ensemble of $2n+1$ base classifier makes the correct decision is $P_c(2n+1)$

$$P_c(2n+1) = \sum_{m=n+1}^{2n+1} \binom{2n+1}{m} p^m (1-p)^{2n+1-m}$$

Question 1 (cont.)

- It can be proved that, with an odd number N
  - If $p > 0.5$ then $P_c(N)$ is monotonically increasing in N, and $P(N) \to 1$ as $N \to \infty$
  - If $p = 0.5$ then $P_c(N) = 0.5$ for all N
  - If $p < 0.5$ then $P_c(N)$ is monotonically decreasing in N, and $P(N) \to 0$ as $N \to \infty$
- Detailed proof can be found in the paper "Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance, 1997"

Table 1: Data set for Question 2.

ID	f_1	f_2	f_3	f_4	f_5	Ground Truth
P1	+	+	-	-	+	+
P2	+	+	-	+	-	+
P3	-	-	+	+	-	+
P4	-	-	+	-	+	+
P5	-	-	+	+	-	-
P6	-	-	-	+	+	+
P7	+	+	+	+	-	+
P8	-	+	+	-	+	-
P9	+	+	-	+	+	+
P10	-	-	-	+	-	-

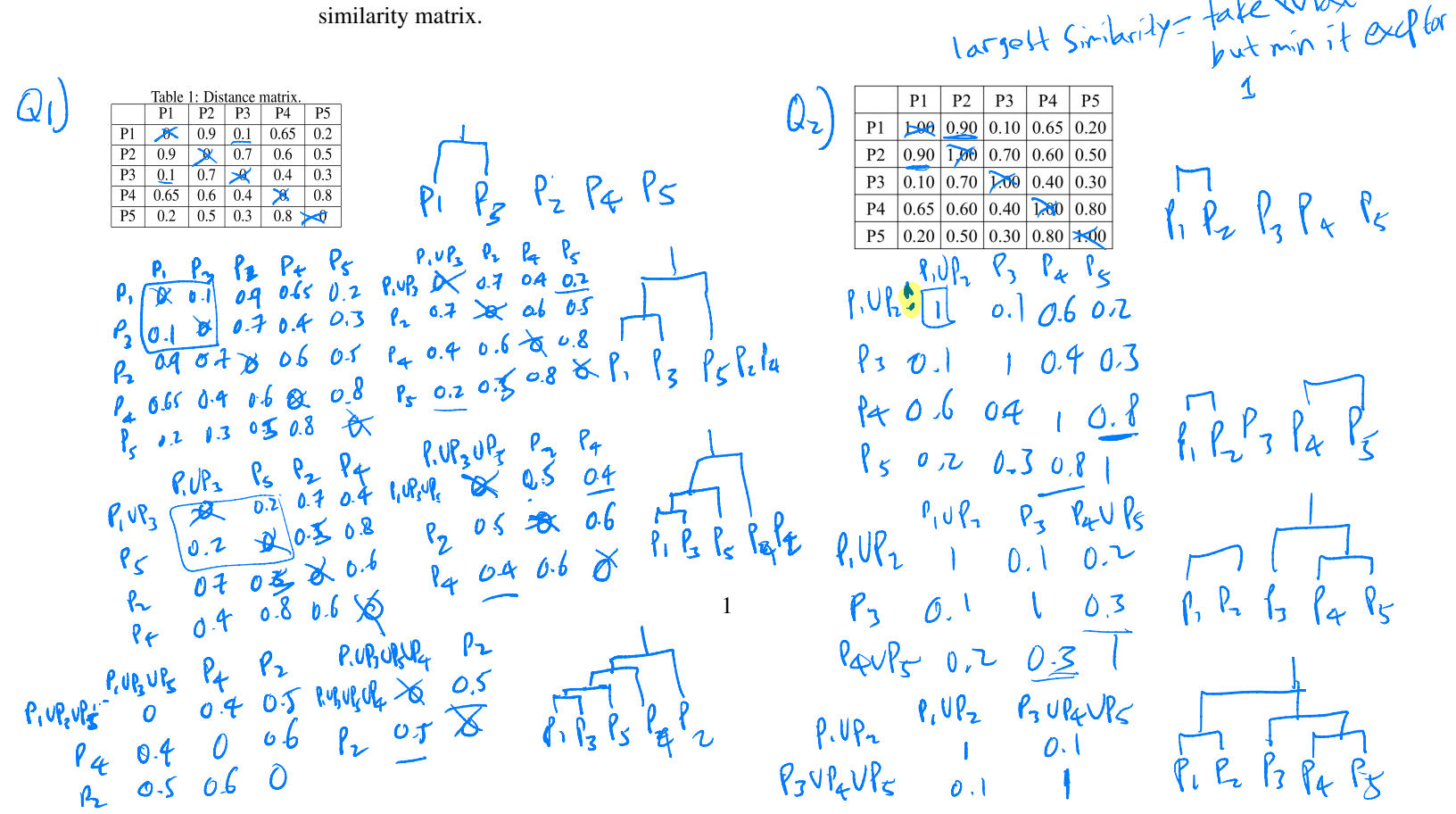
SC4000 Machine Learning Tutorial
Clustering

Question 1: Given the distance matrix shown in Table 1, use a dendrogram to show how to perform agglomerative hierarchical clustering algorithm with Single Link on the distance matrix.

Table 1: Distance matrix.

	P1	P2	P3	P4	P5
P1	0	0.9	0.1	0.65	0.2
P2	0.9	0	0.7	0.6	0.5
P3	0.1	0.7	0	0.4	0.3
P4	0.65	0.6	0.4	0	0.8
P5	0.2	0.5	0.3	0.8	0

Question 2: Refer to the clustering problem on Slide 59 of Lecture 10. Use a dendrogram to show how to perform hierarchical clustering with Complete Link on the similarity matrix.



Question 1

- Suppose a dataset of four 3-dimensional instances is shown below. Estimate the sample mean and covariance matrix (unbiased).

Data matrix

	X_1	X_2	X_3
P1	3	5	−1
P2	−1	8	3
P3	2	−4	−4
P4	0	−1	−6

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i$$

$$\hat{\mu} = \frac{1}{4}\sum_{i=1}^{4} Pi$$

$\hat{\boldsymbol{\mu}}$	X_1	X_2	X_3
	1	2	−2

Centered data matrix

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{i=1}^{N} \tilde{\boldsymbol{x}}_i (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T = \frac{1}{N-1}\sum_{i=1}^{N} \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^T = \frac{1}{N-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$$

$\mathbf{X}$

X_1	X_2	X_3
3	5	−1
−1	8	3
2	−4	−4
0	−1	−6

$\hat{\boldsymbol{\mu}}$	X_1	X_2	X_3
	1	2	−2

$\tilde{\mathbf{X}}$

	X_1	X_2	X_3
	2	3	1
	−2	6	5
	1	−6	−2
	−1	−3	−4

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{4-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \begin{bmatrix} 3.33 & -3 & -2 \\ -3 & 30 & 19 \\ -2 & 19 & 15.33 \end{bmatrix}$$

Question 2

- First, write down the joint distribution function

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i)$$

- Take the logarithm

$$\log P(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \log\lambda - \lambda x_i = n\log\lambda - \sum_{i=1}^{n} \lambda x_i$$

Question 2

- We want to find λ^* that maximizes the joint probability. Therefore, we take the derivative w.r.t λ and set it to zero.

$$\frac{\partial \log P(x_1, x_2, \ldots, x_n)}{\partial\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

$$\frac{1}{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}$$

$$\lambda^* = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\hat{\mu}}$$

$\hat{\mu}$ is the sample mean.

4000 Machine Learning Tutorial
Density Estimation

Question 1: Suppose a dataset of four 3-dimensional instances is shown in Table 1. Estimate the sample mean and covariance matrix (unbiased).

Table 1: Data set for Question 1.

ID	X_1	X_2	X_3
P1	3	5	-1
P2	-1	8	3
P3	2	-4	-4
P4	0	-1	-6

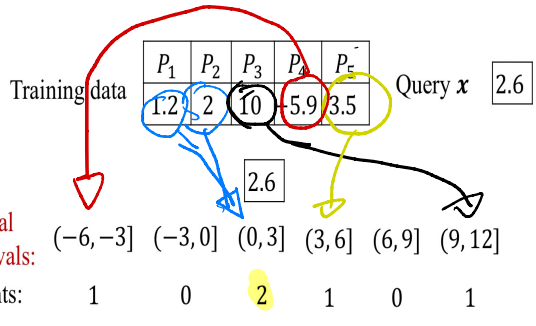
Question 2: The exponential distribution is often used to describe the amount of time between two events (subject to certain assumptions), such as the time between two emails arriving at your inbox, or the time between two people exiting an MRT station. The probability density function is $P(x) = \lambda \exp(-\lambda x)$ with the support $x \in [0, \infty)$. λ is the single parameter that characterizes the distribution. Suppose we have n data points $x_1, x_2, \ldots, x_n$ drawn independently from an exponential distribution. Derive the maximum likelihood estimate for λ.

Question 3: Suppose a dataset of five scalar (1-dimensional) data instances is shown in Table 2. Use histogram estimator with an origin of 0 and a width of 3, naive estimator with a width of 3, and 3-NN estimator to estimate the density function $\hat{P}(\mathbf{x})$ and compute the value of $\hat{P}(2.6)$ at 2.6, respectively.

Table 2: Data set for Question 2.

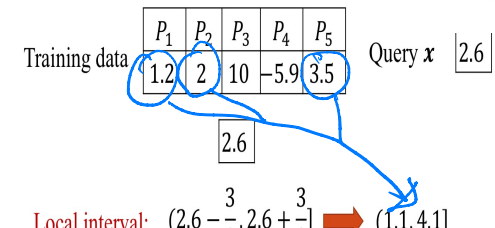
P1	P2	P3	P4	P5
1.2	2	10	-6	3.5

Histogram Estimator (cont.)



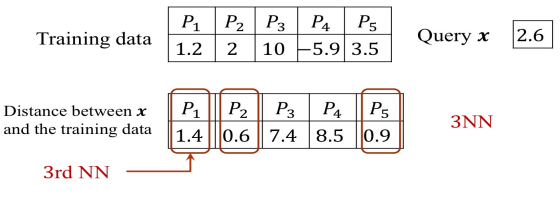
$$\hat{p}(2.6) = \frac{\#\{P_i \mid P_i \text{ in the same bin as } 2.6\}}{N\Delta} = \frac{2}{5 \times 3} = 0.13$$

Naïve Estimator (cont.)



$$\hat{p}(2.6) = \frac{\#\{P_i \mid P_i \text{ in the same bin as } 2.6\}}{N\Delta} = \frac{3}{5 \times 3} = 0.2$$

K-NN Estimator (cont.)



$$\hat{p}(\boldsymbol{x}) = \frac{K}{N2d_K(\boldsymbol{x})} \longrightarrow \hat{p}(2.6) = \frac{3}{5 \times 2 \times 1.4} = 0.21$$

Question 1

- A dataset of five 4-dimensional instances is given in Table 1. Suppose an SVD is performed on the data matrix **X** (5-by-4) via $\mathbf{X} = \mathbf{VDU}^\top$. The matrices **V**, **D**, and **U** are shown in Tables 2-4, respectively. Use principal component analysis to project the 5 data points in Table 1 to 2-dimensional space.

X

X_1	X_2	X_3	X_4
2	4	1	3
1	2	3	5
−2	−4	−4	−1
0	−1	−2	−6
−1	−1	2	−1

Step 1: Center the data points s.t. the mean is **0**

$\hat{\boldsymbol{\mu}}$

X_1	X_2	X_3	X_4
0	0	0	0

Already centered

PCA

- Step 1: Center the data points s.t. the mean is **0**
- Step 2: Compute sample covariance matrix $\tilde{\boldsymbol{\Sigma}}$
- Step 3: Compute eigenvectors of $\tilde{\boldsymbol{\Sigma}}$, $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_d\}$, which are sorted based on their eigenvalues in non-increasing order, i.e., $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$
- Step 4: Select the first k eigenvectors to construct principal components
- Step 5: Project the data instances to k-dimensional space

Dimensionality Reduction

Question 1: A dataset of five 4-dimensional instances is given in Table 1. Suppose an SVD is performed on the data matrix **X** (5-by-4) via $\mathbf{X} = \mathbf{VDU}^\top$. The matrices **V**, **D**, and **U** are shown in Tables 2-4, respectively. Use principal component analysis to project the 5 data points in Table 1 to 2-dimensional space.

Table 1: Data set for Question 1.

Data Points	X_1	X_2	X_3	X_4
P1	2	4	1	3
P2	1	2	3	5
P3	-2	-4	-4	-1
P4	0	-1	-2	-6
P5	-1	-1	2	-1

U = the Eigenvectors

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{5-1}\mathbf{X}^T\mathbf{X} = \frac{1}{4}(\mathbf{VDU}^T)^T\mathbf{VDU}^T$$

$$= \frac{1}{4}\mathbf{UD}^T\underbrace{\mathbf{V}^T\mathbf{V}}_{\mathbf{I}}\mathbf{DU}^T$$

$$\tilde{\mathbf{D}} = \frac{1}{4}\mathbf{D}^T\mathbf{D} \qquad = \mathbf{U}\tilde{\mathbf{D}}\mathbf{U}^T$$

U = the Eigenvectors

The matrix **D** (5-by-4) obtained by SVD ($\mathbf{X} = \mathbf{VDU}^\top$)

10.9040	0	0	0
0	4.8385	0	0
0	0	3.3973	0
0	0	0	0.3867
0	0	0	0

$$\tilde{\boldsymbol{\Sigma}} = \frac{1}{5-1}\mathbf{X}^T\mathbf{X} = \frac{1}{4}(\mathbf{VDU}^T)^T\mathbf{VDU}^T$$

$$= \frac{1}{4}\mathbf{UD}^T\underbrace{\mathbf{V}^T\mathbf{V}}_{\mathbf{I}}\mathbf{DU}^T$$

$$\tilde{\mathbf{D}} = \frac{1}{4}\mathbf{D}^T\mathbf{D} \qquad = \mathbf{U}\tilde{\mathbf{D}}\mathbf{U}^T$$

$$\tilde{\boldsymbol{\Sigma}}\mathbf{U} = \mathbf{U}\tilde{\mathbf{D}}$$

$$\tilde{\boldsymbol{\Sigma}}\boldsymbol{u}_i = \tilde{\mathbf{D}}_{ii}\boldsymbol{u}_i$$

$\mathbf{U}^T\mathbf{U}=\mathbf{I}$ $\qquad \tilde{\mathbf{D}}$

$$\frac{1}{4} \times \begin{bmatrix} 10.90^2 & 0 & 0 & 0 \\ 0 & 4.84^2 & 0 & 0 \\ 0 & 0 & 3.40^2 & 0 \\ 0 & 0 & 0 & 0.39^2 \end{bmatrix}$$

Question 1 (cont.)

$$A = U^T\begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_D \end{bmatrix}U$$

$$A = [u_1 \quad u_2 \quad \dots \quad u_D]\begin{bmatrix} \lambda_1 u_1^T \\ \lambda_2 u_2^T \\ \vdots \\ \lambda_D u_D^T \end{bmatrix}$$

$$\alpha A = \alpha\lambda_1 x_1 x_1^T + \alpha\lambda_1 x_2 x_2^T + \cdots + \alpha\lambda_D x_D x_D^T$$

$$A = [u_1 \quad u_2 \quad \dots \quad u_D]\begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_D \end{bmatrix}\begin{bmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_D^T \end{bmatrix}$$

$$A = \lambda_1 u_1 u_1^T + \lambda_2 u_2 u_2^T + \cdots + \lambda_D u_D u_D^T$$

$$\alpha A = \alpha\lambda_1 u_1 u_1^T + \alpha\lambda_1 u_2 u_2^T + \cdots + \alpha\lambda_D u_D u_D^T$$

$$\alpha A = U^T\begin{bmatrix} \alpha\lambda_1 & 0 & 0 & 0 \\ 0 & \alpha\lambda_2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \alpha\lambda_D \end{bmatrix}U$$

Question 1 (cont.)

- Step 2: Compute sample covariance matrix $\widetilde{\boldsymbol{\Sigma}}$
- Step 3: Compute eigenvectors of $\widetilde{\boldsymbol{\Sigma}}$, $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d\}$, which are sorted based on their eigenvalues in non-increasing order, i.e., $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$

- Perform SVD on the centered data matrix $\mathbf{X}$ (5-by-4 to obtain
  $\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{U}^\top$ (already done in this question)

- Step 4: Select the first k eigenvectors to construct principal components

- Select the first 2 eigenvectors of $\widetilde{\boldsymbol{\Sigma}}$, i.e, the first two column vectors of the matrix $\mathbf{U}$ to construct the principle component matrix $\mathbf{U}_2 = [\boldsymbol{u}_1, \boldsymbol{u}_2]$

$\mathbf{U}$ (with $\mathbf{U}_2$ highlighted):

−0.2224	0.3430	−0.3302	−0.8508
−0.4880	0.5756	−0.4046	0.5166
−0.4479	0.2924	0.8400	−0.0911
−0.7154	−0.6823	−0.1473	−0.00309

Table 2: The matrix $\mathbf{V}$ (5-by-5) obtained by SVD ($\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{U}^\top$)

-0.4577	0.2550	-0.5536	0.4680	0.4472
-0.5612	-0.2150	0.1896	-0.6348	0.4472
0.4497	-0.7183	-0.2749	0.0787	0.4472
0.5206	0.6063	-0.1153	-0.3849	0.4472
0.0486	0.0720	0.7542	0.4730	0.4472

Table 3: The matrix $\mathbf{D}$ (5-by-4) obtained by SVD ($\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{U}^\top$)

10.9040	0	0	0
0	4.8385	0	0
0	0	3.3973	0
0	0	0	0.3867
0	0	0	0

Table 4: The matrix $\mathbf{U}$ (4-by-4) obtained by SVD ($\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{U}^\top$)

-0.2224	0.3430	-0.3302	-0.8508
-0.4880	0.5756	-0.4046	0.5166
-0.4479	0.2924	0.8400	-0.0911
-0.7154	-0.6823	-0.1473	-0.0309

Question 1 (cont.)

- Step 5: Project the instances to k-dimensional space

- Compute $\mathbf{X}\mathbf{U}_2$

2	4	1	3
1	2	3	5
−2	−4	−4	−1
0	−1	−2	−6
−1	−1	2	−1

$\mathbf{X}$ ×

−0.2224	0.3430
−0.4880	0.5756
−0.4479	0.2924
−0.7154	−0.6823

$\mathbf{U}_2$ =

−4.9908	1.2338
−6.1191	−1.0402
4.9038	−3.4759
5.6762	2.9335
0.5399	0.3486

$\mathbf{Z}$