Recitation 6

Exam 1 Review

10-701: Introduction to Machine Learning

10/27/2023

1 Decision Trees

1. You are given a dataset for binary classification with two features x_1 and x_2. x_1 can have two possible values 0,1 and x_2 can have three possible values 0,1,2. Figure 1 provides a depiction of this dataset.

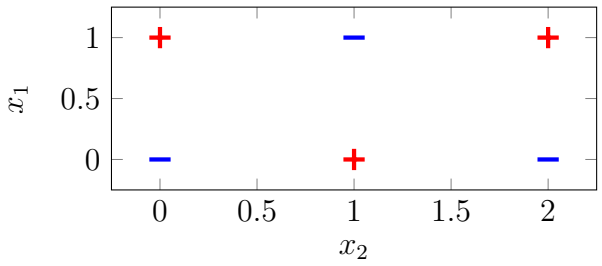


Figure 1: Binary classification dataset.

   (a) What is the lowest possible training error any decision tree could achieve on this dataset? The tree need not be a binary tree but each node should split on at most one feature.

   0

   (b) Draw a decision tree that achieves this training error.

Many possible solutions exist but the most intuitive one first split on $x_1 < 0.5$ and then makes a ternary split on x_2 down either branch at 0.5 and 1.5.

2. We would like to learn a decision tree. We have n samples for training. You can assume that n is large and we are using continuous features. In the following questions, suppose we only use the first feature, x_1.

   (a) We would like to split according to x_1 at the root with 3 branches (samples are split at the root to three different sub-trees which by findings values a and b such that the three sub-trees are $x_1 \leq a$, $a < x_1 < b$ and $x_1 \geq b$). What is the runtime for finding the values of x_1 that should be used by the root for such a split?

   - ◯ $O(\log n)$
   - ◯ $O(\sqrt{n})$
   - ◯ $O(n)$
   - ◯ $O(n^2)$
   - ◯ $O(n^3)$

   D. To split on continues values we need to find the best thresholds a and b. This requires us to order the values of x_1 and then test all possible splits and so the runtime is $O(n^2)$.

   (b) Following our first split, we would like to *split again* on x_1 in each of the three sub-trees resulting from the split in the previous question. Again, for each sub-tree we would split three ways as we did in the root. What is the total run-time for determining the optimal splits for ALL three sub-trees?

   - ◯ $O(n)$
   - ◯ $O(n^2)$
   - ◯ $O(n^3)$
   - ◯ $O(n^4)$
   - ◯ $O(n^8)$

   B. Let the number of samples assigned to each of the three sub-trees be n_1, n_2 and n_3. We know from question 1 that the total run time for each is $O(n_1^2)$, $O(n_2^2)$, $O(n_3^2)$. Since $n_1 + n_2 + n_3 = n$ one of these sets is $O(n)$ and so the total runtime is $O(n^2)$.

(c) For the same dataset we would like to learn the *best* tree that:

- Only uses x_1
- Has two levels, a root and a level below it where each splits to three branches (see the figure below).

What is the runtime for computing the *optimal tree*, in terms of minimizing the training error?

- ◯ $O(n)$
- ◯ $O(n^2)$
- ◯ $O(n^4)$
- ◯ $O(n^8)$
- ◯ $O(n^9)$

D. Finding the best way to split in this case is similar to finding the best 9 way split. This means that the total run time is $O(n^8)$.

2 kNN

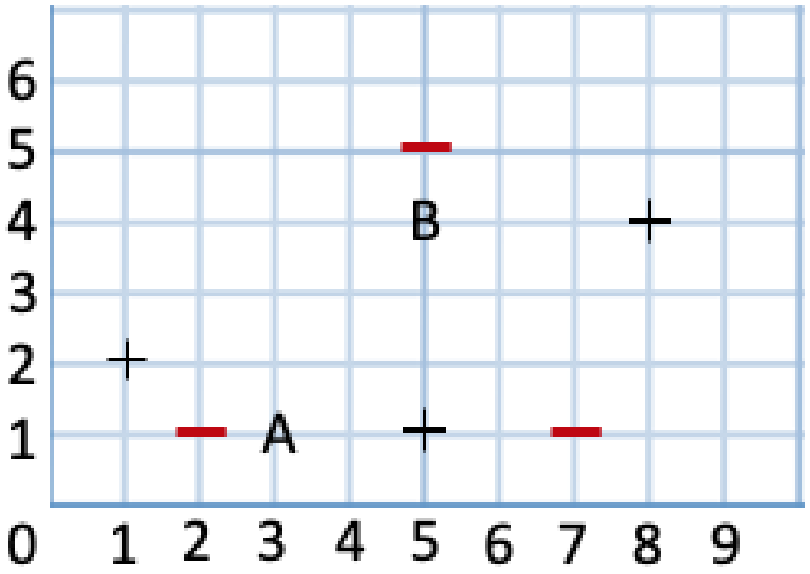
1. Imagine you are deciding between the following kNN models for a binary classification problem where $y \in \{-1, +1\}$ (for both models, assume we use the Euclidean distance and *the nearest neighbor for a training data point is the point itself*):

   - Model 1: Probabilistic 3NN. This model predicts the label of a query data point with probabilities corresponding to the relative proportion of labels in the 3 nearest neighbors, e.g., if the 3 nearest neighbors have labels $\{-1, -1, +1\}$ then this model predicts -1 with probability $2/3$ and $+1$ with probability $1/3$.
   - Model 2: Distance weighted 2NN. This model predicts

   $$\hat{y} = \text{sign}\left(\sum_{(\vec{x}^{(i)}, y^{(i)}) \in N_2(\vec{x})} \frac{y^{(i)}}{\|\vec{x}^{(i)} - \vec{x}\|_2^2 + 1}\right)$$

   where $N_2(\vec{x})$ are the 2 nearest neighbors in the training data to the query data point $\vec{x}$ (the plus one in the denominator is to avoid division by 0). For the purposes of this model, let $\text{sign}(0) = +1$.

   You gather a training dataset of 6 points: 3 red $-$'s, which correspond to label $y = -1$, and 3 black $+$'s, which correspond to label $y = +1$. You also gather a validation dataset consisting of points A and B, both of which have label $+1$.



(a) What is the *expected* training error rate for Model 1? You may express your answer as a fraction if necessary.

Starting with the -1 at (5,5) and going clockwise, the expected pointwise training errors are

$$\frac{2}{3}+\frac{2}{3}+\frac{2}{3}+\frac{2}{3}+\frac{2}{3}+\frac{1}{3}=\frac{11}{3}$$

Thus, the expected training error rate is $\frac{11}{18}$.

(b) What is the training error rate for Model 2? You may express your answer as a fraction if necessary.

0; the trick here is to realize that this is effectively a 1NN model when applied to the training data as the nearest neighbor to a training data point will always be itself and thus, will always have a higher weight than the second nearest neighbor.

(c) What is the *expected* validation error rate for Model 1? You may express your answer as a fraction if necessary.

$\frac{1}{3}$; the expected validation error on both A and B is $\frac{1}{3}$ as the 3 nearest neighbors to A and B both have label sets $\{-1,+1,+1\}$.

(d) What is the validation error rate for Model 2? You may express your answer as a fraction if necessary.

1; again, because no two points are equidistant to either A or B, this model boils down to a 1NN.

2. For the points shown in Figure 2 suppose x_1, x_4, x_5, x_7, and x_9 have the label 1, and the other points have the label 0. For a 3-NN classifier using the Euclidean distance, what is the LOOCV error?

The predictions are below. The error is 1

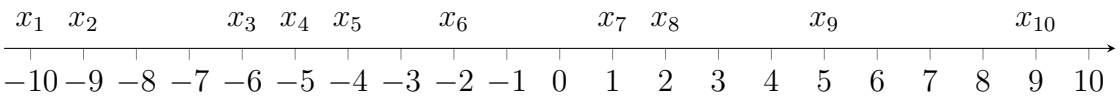


Figure 2

point	ground truth	LOOCV prediction
x_1	1	0
x_2	0	1
x_3	0	1
x_4	1	0
x_5	1	0
x_6	0	1
x_7	1	0
x_8	0	1
x_9	1	0
x_{10}	0	1

3 Linear Regression & Regularization

1. Using monthly stock averages, $x_1, x_2, \ldots x_n$, you run linear regression without regularization to estimate some future stock value y. However, your test error turns out to be very high. Your friend suggests that maybe the price *differences* between consecutive months may offer better features, so you add additional features $< (x_1 - x_2), (x_2 - x_3), (x_3 - x_4), ..., (x_{n-1} - x_n) >$. Do you expect an improvement? Justify your answer in 2-3 concise sentences.

Solution: No improvement as the features you are using are a linear combination of features you already have. So for example if you have:

$$w_1 x_1 + w_2 x_2 = y$$

In our original model in our new model:

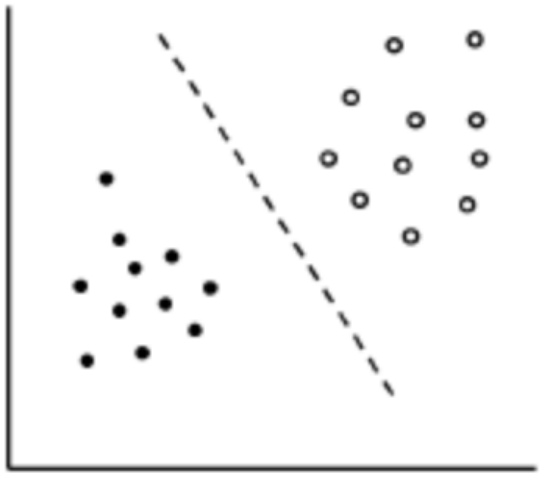
$$w_1 x_1 + w_2 x_2 + w_3 (x_1 - x_2) = y$$

Then this is equivalent to:

$$w_1 x_1 + w_2 x_2 + w_3 x_1 - w_3 x_2 = (w_1 + w_3) x_1 + (w_2 - w_3) x_2$$

Then these weights will balance out to be equivalent to the original w_1 and w_2.

2. The plot below shows data from two classes (filled and unfilled circles) separated by a linear decision boundary (dashed) defined by a two-dimensional weight vector and a bias term.



Suppose we apply L_2 regularization *to the bias term only* i.e., the two-dimensional weight vector remains the same for the regularized and unregularized settings. Which of the following best describes the new decision boundary that would result from this change. Assume that the amount of regularization is mild.

- ◯ farther from the origin; parallel to the previous line
- ◯ farther from the origin; not parallel to the previous line
- ◯ closer to the origin; parallel to the previous line
- ◯ closer to the origin; not parallel to the previous line

C

4 MLE/MAP & Naïve Bayes

1. We first define the Pareto distribution as:

$$p(x|k,\alpha) = \begin{cases} \frac{\alpha k^\alpha}{x^{\alpha+1}} & x \in [k,\infty) \\ 0 & \text{otherwise} \end{cases}$$
$$k, \alpha \in (0,\infty)$$

Given n independent samples $x_1, x_2, ..., x_n$ drawn from a Pareto distribution, what is the MLE for the parameters k and α? **Hint: first determine the MLE of k, then use that result to in your expression for the MLE of α.**

$$\hat{k}_{MLE} = \_\_\_\_\_$$

$$\hat{\alpha}_{MLE} = \_\_\_\_\_$$

$$\hat{k}_{MLE} = \min_i x_i$$
$$\hat{\alpha}_{MLE} = \frac{n}{\sum_{i=1}^n \ln\left(\frac{x_i}{\hat{k}_{MLE}}\right)}$$

2. Suppose we are training a Gaussian Naïve Bayes classifier to distinguish between students taking 10701 and 10315 based on 12 real-valued features. We train two models: M_1 using data from students across both courses in the F23 semester, and M_2 using values from the S23 semester. For F23, we had G_1 students in 10701, and U_1 students from 10315. For S23, we had G_2 and U_2 students respectively.

   Amazingly, M_1 and M_2 learned the same values for all the means and variances!

   Suppose a new student which we did not use in training shows up and we classify them using M_1 and M_2. M_1 predicts they are in 10701 whereas M_2 predicts they are in 10601. Given this, which of the following relationships must be true?

   - ◯ $G_1 > G_2$
   - ◯ $U_2 > U_1$
   - ◯ $G_1 + U_1 > G_2 + U_2$

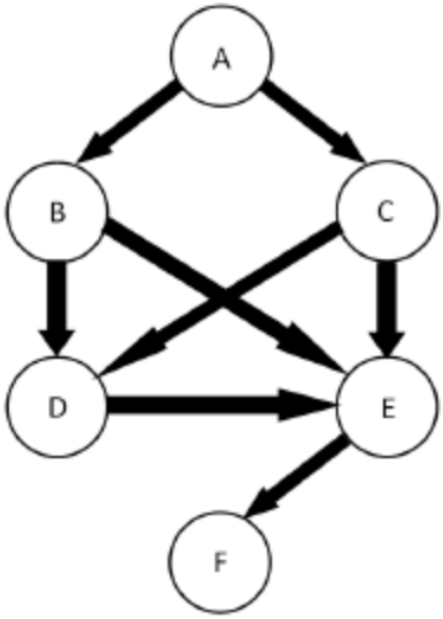
○ $\frac{G_1}{G_2} > \frac{U_1}{U_2}$

Answer: D. For this result to happen we need the prior for 10701 to be higher in M1 than it is in M2. This means that $\frac{G1}{G1+U1} > \frac{G2}{G2+U2}$. All values in this inequality are positive so this is the same as:

$$\frac{G2+U2}{G2} > \frac{G1+U1}{G1} \implies \frac{U2}{G2}+1 > \frac{U1}{G1}+1 \implies \frac{G1}{G2} > \frac{U1}{U2}$$

5 Bayesian Networks & Causality

1. Suppose we have a graphical model below made of random variables A, B, C, D, E and F, where B and C can take on values $\{0, 1, 2\}$ and all of the other variables are binary.



(a) What is the joint distribution of this graphical model?

$P(A, B, C, D, E, F) =$ $= P(A)P(B|A)P(C|A)P(D|B, C)P(E|B, C, D)P(F|E)$

(b) How many parameters do we need to store for this graphical model?

$P(A) = 1, P(B|A) = 4, P(C|A) = 4, P(D|B, C) = 9, P(E|B, C, D) = 18, P(F|E) = 2$
So 38

(c) For the following independences please give the smallest set of variables we need to condition on to make the statement true. *The empty set is a valid choice.*

i. $A \perp E|$

ii. $B \perp C|$

iii. $B \perp F|$

i. C,B
ii. A
iii. E

2. Researchers want to study whether taking a certain type of vitamin helps a patient sleep. Consider the table below: let X indicate whether a person takes the vitamin, Y indicate whether the person sleeps well, P_1 indicate whether the person sleeps well if they start taking the vitamin, and P_2 indicate whether the person sleeps well if they stop taking the vitamin.

X	Y	P_1	P_2
0	0	0	0
0	0	0	0
1	1	1	0
0	0	0	1
0	0	1	1
1	1	0	1
1	1	1	1
1	1	1	0

Table 1: Causal Table

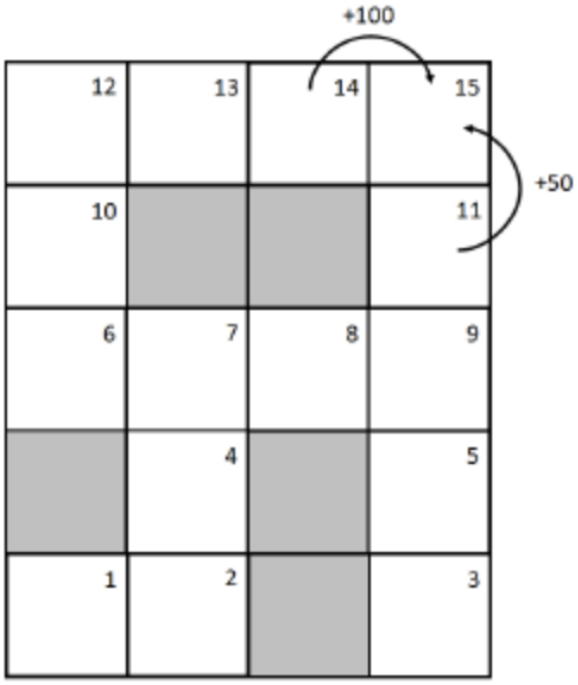
Calculate the correlation and average treatment effect between taking the vitamin and having good sleep quality. Does correlation indicate causality in this case?

Correlation: 1, Casual effect 0 Correlation does not indicate causality

6 Reinforcement Learning

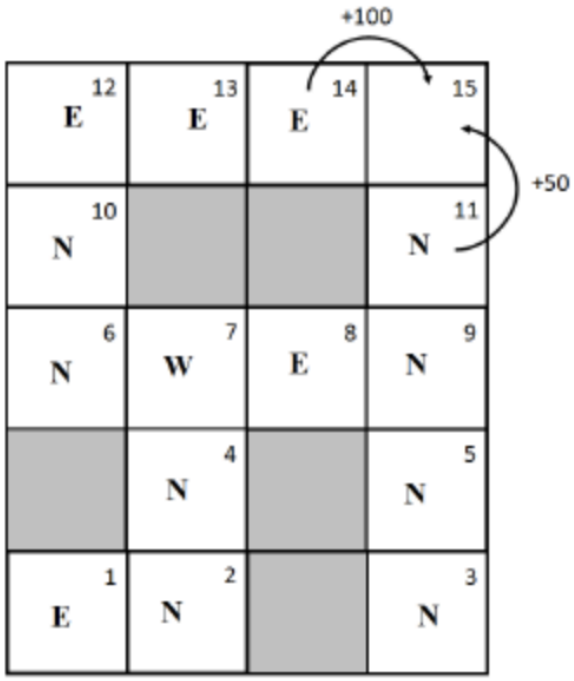
1. Your agent is tasked with navigating the maze shown below. The states are labelled from 1 - 15, with 15 being a terminal state. Boxes in grey cannot be accessed by the agent. Valid actions are denoted by North (↑), East (→), South (↓) and West (←). The initial discount factor is $\gamma = 0.8$ and the rewards for this network are as follows:

   - +50 Reward for moving from state 11 to 15.
   - +100 Reward for moving from state 14 to 15.
   - +0 All other state transitions.



   (a) On the maze above, draw the optimal policy π^* that the agent would take.

   SOLUTION:



(b) What is the value of $Q(S_{12}, (\rightarrow))$?

$0.8^2 * 100 = 64$

(c) What is the value of $Q(S_{12}, (\downarrow))$?

$0.8^4 * 100 = 40.96$

(d) For what value of γ are the actions $\rightarrow$ and $\leftarrow$ equally good in S_7?

This can be calculated in the following way: $\gamma^5 100 = \gamma^3 50 \implies \gamma^2 = 1/2 \implies \gamma = 1/\sqrt{2}$

(e) **Select all that apply:** Which of the following changes to our maze game could make the optimal policy starting from S_3 get reward: $R(S_{14}, \rightarrow)$?

- ◯ Changing $R(S_{14}, \rightarrow) = 200$
- ◯ Changing $R(S_{11}, \uparrow) = 30$
- ◯ Changing the transitions so that they are not deterministic
- ◯ Changing $\gamma = 1/\sqrt[5]{2}$

A, C

2. Suppose you are training an agent with Q-learning in the following maze environment over states $\mathcal{S} = \{A, B, C, D\}$ with actions $\mathcal{A} = \{up, down, left, right\}$. The rewards $R(s, a)$ for $s \in \mathcal{S}$ and $a \in \mathcal{A}$ are shown on the arrows between states. The arrows also show the allowable transitions between states. Assume below a discount factor of $\gamma = 0.9$. We initialize $V(s) = 0$ for all state s. State D is terminal.

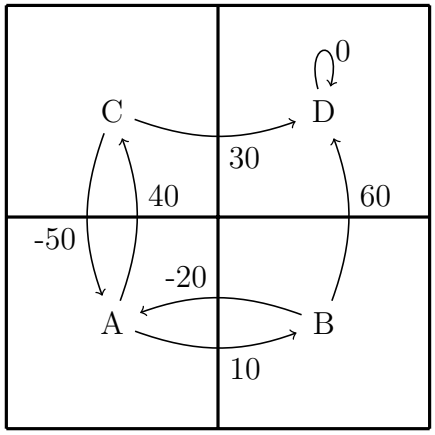


Figure 3

(a) What is the value $V^*(A)$? [ ]

$$\begin{aligned} V^*(A) &= max(r_{up}, r_{right}) \\ &= max(40 + 0.9 * 30, 10 + 0.9 * 60) \\ &= max(67, 64) \\ &= 67 \end{aligned} \tag{1}$$

(b) What is the value $Q^*(A, right)$? [ ]
64, refer to calculations in part (a)

(c) **Select all that apply:** What action does the optimal policy take from state A (i.e. what is $\pi^*(A)$)? (Note: If the optimal policy is not unique and there are multiple optimal actions, select them all.)

☐ up

☐ down

☐ left

☐ right

Up only

7 Neural Networks

1. Your friend wants to build a fully-connected one-hidden-layer Neural Network. Her inputs have 99 features. She wants 7 hidden neurons in the first layer and 2 neurons in the output layer generated using a softmax function. Each layer is a linear layer fully-connected to the previous one and includes a bias term. Each scalar counts as one parameter.

   (a) What is the number of parameters she will need to create this neural network? Each scalar counts as one parameter.

   (99+1)x7 for the first layer is 700, (7+1)x2 for the output layer is 16, so the total is 716.

   (b) Your friend shares with you that the output is binary. Using this information is there a way to learn the same decision boundary and reduce the number of parameters? Justify your answer in 1-2 concise sentences

   Yes, she could make the output layer be only one node which contains the sigmoid function. Using a threshold of 0.5 will then recreate the predictions of the softmax.

2. Your friend trains a Neural Network on some training dataset and achieves an almost perfect training accuracy. However they perform very poorly when it comes to test time. In the boxes below, write either **{Increase, Decrease or N/A}** to indicate what they should do with their model to reduce the amount of overfitting (N/A here means changing the specified quantity won't effect overfitting). *You should assume the training and test data are representative of the true underlying distribution of the data.*

   the number of training data

   the number of test data

   the regularization constant

   the number of hidden layers

   the number of neurons in each hidden layer

   the number of nodes in the output layer

Increase
N/A
Increase
Decrease
Decrease
N/A