Machine Learning Homework 1

Zihan Shao

Instructions

- **Online submission:** You must submit your solutions online on NYU Classes. We recommend that you use LaTeX, but we will accept Word document, scanned / pictured solutions as well.
- **Submission format** Please submit your homework with a single zip file. Files in the zip file are
  - NetID-HW1.pdf : a write-up file which contains answers to problems, and it should contain the commands for running your code;
  - problem-X.py : the code files for problems (where X is the ID of the problem).

Problem 1: More Probability Review

(a) [**5 Points**] For events A and B, prove

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Solution: Rewrite $P(B|A)$ by the definition of Conditional probability,

$$\text{Right Hand Side} = \frac{P(B|A)P(A)}{P(B)} = \frac{\frac{P(A,B)}{P(A)} \cdot P(A)}{P(B)} = \frac{P(A,B)}{P(B)} = P(A|B) = \text{Left Hand Side}$$

(b) [**5 Points**] For events A, B, and C, rewrite $P(A, B, C)$ as a *product* of several conditional probabilities and one unconditional probability involving a single event. Your conditional probabilities can use only one event on the left side of the conditioning bar. For example, $P(A|C)$ and $P(A)$ would be okay, but $P(A, B|C)$ is not.

Solution: $P(A, B, C) = P(A) \cdot P(B|A) \cdot P(C|A, B)$. The position of A, B, and C can be exchanged without loss of generality.

(c) [**5 Points**] Let A be any event, and let X be a random variable defined by

$$X = \begin{cases} 1 & \text{if event } A \text{ occurs} \\ 0 & \text{otherwise.} \end{cases}$$

X is sometimes called the indicator random variable for the event A. Show that $\mathbb{E}[X] = P(A)$, where $\mathbb{E}[X]$ denotes the *expected value* of X.

Solution:

$$\mathbb{E}[X] = 1 \cdot P(A) + 0 \cdot P(A^c) = P(A)$$

(d) Let X, Y, and Z be random variables taking values in $\{0, 1\}$. The following table lists the probability of each possible assignment of 0 and 1 to the variables X, Y, and Z:

	$Z = 0$		$Z = 1$	
	$X = 0$	$X = 1$	$X = 0$	$X = 1$
$Y = 0$	1/15	1/15	4/15	2/15
$Y = 1$	1/10	1/10	8/45	4/45

For example, $P(X = 0, Y = 1, Z = 0) = 1/10$ and $P(X = 1, Y = 1, Z = 1) = 4/45$.

(i) **[5 Points]** Is X independent of Y? Why or why not?

Solution: No. According to the definition of Independence of Random Variables, discrete random variables X and Y are independent if the events $X = x$ and $Y = y$ are independent for all x and y. In this case:

$$P(X = 0) = 1/15 + 1/10 + 4/15 + 8/45 = 11/18$$

$$P(Y = 0) = 1/15 + 1/15 + 4/15 + 2/15 = 8/15$$

However, $P(X = 0, Y = 0) = 1/15 + 4/15 = 1/3 \neq 44/135 = P(X = 0) \cdot P(Y = 0)$, which implies that events $X = 0$ and $Y = 0$ are not independent. It follows that X and Y are not independent.

(ii) **[5 Points]** Is X conditionally independent of Y given Z? Why or why not?

Solution: Yes. Similarly, we need to verify if $P(X, Y|Z) = P(X|Z) \cdot P(Y|Z)$. By the information provided, we have:

$$P(X = 0|Z = 0) = 1/2 \tag{1}$$
$$P(X = 1|Z = 0) = 1/2 \tag{2}$$
$$P(Y = 0|Z = 0) = 2/5 \tag{3}$$
$$P(Y = 1|Z = 0) = 3/5 \tag{4}$$
$$P(X = 0|Z = 1) = 2/3 \tag{5}$$
$$P(X = 1|Z = 1) = 1/3 \tag{6}$$
$$P(Y = 0|Z = 1) = 3/5 \tag{7}$$
$$P(Y = 1|Z = 1) = 2/5 \tag{8}$$
$$P(X = 0, Y = 0|Z = 0) = 1/5 \tag{9}$$
$$P(X = 1, Y = 0|Z = 0) = 1/5 \tag{10}$$
$$P(X = 0, Y = 1|Z = 0) = 3/10 \tag{11}$$
$$P(X = 1, Y = 1|Z = 0) = 3/10 \tag{12}$$
$$P(X = 0, Y = 0|Z = 1) = 2/5 \tag{13}$$
$$P(X = 1, Y = 0|Z = 1) = 1/5 \tag{14}$$
$$P(X = 0, Y = 1|Z = 1) = 4/15 \tag{15}$$
$$P(X = 1, Y = 1|Z = 1) = 2/15 \tag{16}$$

Here we can see $P(X, Y|Z) = P(X|Z) \cdot P(Y|Z)$ holds for all the cases, which indicates the independence.

(iii) **[5 Points]** Calculate $P(X = 0|X + Y > 0)$.

Solution:

$$
\begin{aligned}
P(X=0|X+Y>0) &= \frac{P(X=0, X+Y>0)}{P(X+Y>0)} && (17)\\
&= \frac{P(X=0, Y=1)}{1-P(X=0, Y=0)} && (18)\\
&= \frac{1/10+8/45}{1-1/15-4/15} && (19)\\
&= 5/12 && (20)
\end{aligned}
$$

Problem 2: KNN for Iris flowers classification

For this problem, we want to use K nearest neighbor algorithm to classify Iris flowers.

(a) **[5 Points]** Load the Iris data using `sklearn.datasets`. Calculate how many elements there are for every class.

Solution:
There are 50 elements in class target = 0
There are 50 elements in class target = 1
There are 50 elements in class target = 2

(b) **[10 Points]** Build a KNeighborsClassifier with $k = 1$ to predict the class. Train it on the whole dataset. For classification problem, different goodness-of-fit metrics are used. For this exercise, you can use accuracy, defined in the formula given below. Calculate the accuracy of the KNN classifier on the iris dataset. Is it meaningful?

$$\text{Accuracy} = \frac{\#\text{correctly predicted}}{M}$$

Solution: The accuracy is 100%. This result is meaningless since it is always 100%. When we put the training data into the trained model, the k-nearest neighbour, 1-st nearest neighbour in this case, is always itself. In that case, all the elements will be correctly predicted.

(c) **[20 Points]** Split the dataset into two parts using sklearn's `train_test_split`. Use the following arguments:

- X, y : the dataset
- test_size = 0.5 (use 50% of the dataset for testing)
- shuffle = True (randomly shuffle the dataset before making a cut)
- random_state = 0 (random seed, this ensures consistent results)

Use the split to find the optimal value of k. Please try different different value of k from 1 to 50, train the model on the training data and calculate the model's accuracy on the training data and testing data respectively. Plot the training accuracy and testing accuracy against the value of k. Which k value would say is best?

Solution: $K = 9$ is optimal since the testing accuracy is at the peak. (See code file)

(d) **[5 Points]** You observed a flower and measured the following characteristics:

- sepal width of $x_0 = 5.0$
- petal width of $x_1 = 4.1$

- sepal length of $x_2 = 3.8$
- petal length of $x_3 = 1.2$

Use your prediction model to classify this plant. What's the predicted class?

Solution: The predicted class is target = 0, which is **setosa**. (See code file)

Problem 3: K Means clustering

For this question, use the data found in `clust_data.csv`. We will attempt to cluster this data using k-means. But, what k should we use?

(a) [**10 Points**] Apply k-means to this data 15 times, using number of centers from 1 to 15. Each time use nstart = 10 and store the within-cluster sum-of-squares/inertia value from the resulting object. The inertia measures how variable the observations are within a cluster, which we would like to be low. So obviously this value will be lower with more centers, no matter how many clusters there truly are. Plot this value against the number of centers. Look for an "elbow", the number of centers where the improvement suddenly drops off. Based on this plot, how many cluster do you think should be used for this data?

Solution: According to the image we just plotted, $k = 4$ is at the elbow, which suggests that it should be the optimal k. (See code file)

(b) [**10 Points**] Re-apply k-means for your chosen number of centers. How many observations are placed in each cluster? What is the value of inertia?

Solution:
There are 25 observations in the #0 cluster.
There are 25 observations in the #1 cluster.
There are 25 observations in the #2 cluster.
There are 25 observations in the #3 cluster.
The value of inertia is 4844.9258176238245. (See code file)

(c) [**10 Points**]Visualize this data. Plot the data using the first two variables and color the points according to the k-means clustering. Based on this plot, do you think you made a good choice for the number of centers? (Briefly explain.)

Solution: It is not a good choice for the first two dimension.
According to the image, we cannot see 4 clear clusters. The inertia is too big and the inter-cluster variance is too small. This is because the model is based on all 50 dimensions of data. The model does not guarantee the first two features of points in the same cluster are similar.
In some senses, it is the curse of dimensions. The points are clustered well for 50 dimensions but not the first two dimensions. (See code file)