

CMPT 726 Assignment 2

Dhruv Patel, 301471961

1. Cross Validation

a) In order to get an unbiased estimate of how well our ML models and algorithms perform, it is typical to do a 60/20/20 split for our data: 60% train, 20% test and 20% validation. Several years ago this was widely considered best practice in machine learning. Do you still agree such ratios in the modern big data era? Why or why not? (Hint: consider different scenarios)

Answer: Clearly, in this modern big data era. We cannot fix any ratios to split into train, validation, and test dataset as we did several years ago. The decision to split the dataset should be done after considering various scenarios such as

- Availability of data
 - If we have more than enough data available to train the model. Then we can split the data into train, validation and test data.
- Representativeness of data
 - Suitable representation of the scenario, means data mostly covers all the possible cases. Else, we need to make sure. Training and test data have almost similar records in distribution.
- Computational cost on the data
 - Since there will always be a computational cost associated to train a model. Based on the model techniques and the size of the data. To optimise the cost, accordingly we can increase or decrease training data.

As a result, splitting of the records mainly depends on the data and the way to model the problem. Some best estimations considering the fact that we have more than enough data which represents similar scenarios. We can always go for 70% (Training data), 15% (Validation data) and 15% (Test data).

If some data has more noise, in that case we can neglect validation dataset, and given more data to training data (75%) and test data (25%)

b) Consider the hypothetical graph below of predictive error (y-axis) vs. model complexity (x-axis), and how test/training error varies as model complexity increases.

i) Which part of the plot means that the model is Overfitting on the data? (Choose A, B, or C)

Answer: After analysing the given hypothetical graph, we can say **part C** denotes that model is overfitted on data. Since, the gap between error on training set and test set is maximum. Because, model parameters were overly trained on training data so we have low training error and high error on unknown data.

ii) Which part of the plot means that the model is Underfitting on the data? (Choose A, B, or C)

Answer: Based on the hypothetical graph, **part A** denotes that model is under-fitted on data. Both training error and test error are high in that part. Since model capacity is less.

iii) Which part of the plot representing the ideal model complexity? (Choose A, B, or C)

Answer: **Part B** represents the ideal model complexity in the given hypothetical graph. Showing the ideal and best scenario where the gap between both the training and test error is minimum. And both do not deviate that much.

c) For a decision tree model, whose train/test errors are in region A, which of the following is most likely to improve the model performance on real data? (Choose one)

i) Acquire more training data.

ii) Reduce the depth of the decision tree.

iii) Increase the depth of the decision tree.

Answer:

The problem with the decision tree model in region A, is that it's overfitted. So ideally option " i) Acquire more training data would not be helpful ". Even option " iii) Increase the depth of the decision " would make it even worse by more categorising the data it will further train the model. Finally **ii) Reduce the depth of the decision tree** or Pruning is the best way to handle overfitting of a decision tree model.

3. Probabilistic Modeling and Bayes' Rule

a) Assume the probability of being infected with Malaria disease is 0.01. The probability of test positive given that a person is infected with Malaria is 0.95 and the probability of test positive given the person is not infected with Malaria is 0.05.

Answer:

Given,

$$P(\text{malaria}) = 0.01 \text{ and so } P(\text{no malaria}) = 1 - P(\text{malaria}) = 1 - 0.01 = 0.99$$

$$P(\text{positive} \mid \text{malaria}) = 0.95$$

$$P(\text{positive} \mid \text{no malaria}) = 0.05$$

(a) Calculate the probability of test positive.

$$\begin{aligned} P(\text{positive}) &= P(\text{malaria}) * P(\text{positive} \mid \text{malaria}) + P(\text{no malaria}) * P(\text{positive} \mid \text{no malaria}) \\ &= (0.01 * 0.95) + (0.99 * 0.05) \\ &= 0.0095 + 0.0495 \\ &= 0.059 \end{aligned}$$

(b) Use Bayes' Rule to calculate the probability of being infected with Malaria given that the test is positive.

$$\begin{aligned} P(\text{malaria} \mid \text{positive}) &= P(\text{malaria}, \text{positive}) / P(\text{positive}) \\ &= P(\text{malaria}) * P(\text{positive} \mid \text{malaria}) / P(\text{positive}) \\ &= (0.01 * 0.95) / 0.059 \\ &= 0.16101694915 \end{aligned}$$

b) Suppose $P(\text{rain today}) = 0.30$, $P(\text{rain tomorrow}) = 0.60$, $P(\text{rain today and tomorrow}) = 0.25$. Given that it rains today, what is the probability it will rain tomorrow?

Answer:

Given,

$$P(\text{rain today}) = 0.30$$

$$P(\text{rain tomorrow}) = 0.60$$

$$P(\text{rain today} \cap \text{tomorrow}) = 0.25$$

$$\begin{aligned} \text{So, } P(\text{rain tomorrow} \mid \text{rain today}) &= P(\text{rain today} \cap \text{rain tomorrow}) / P(\text{rain today}) \\ &= (0.25 / 0.30) \\ &= 0.833 \end{aligned}$$

c) A biased die has the following probabilities of landing on each face:

Face	1	2	3	4	5	6
P(face)	0.2	0.2	0.1	0.1	0.1	0.3

i) I win if the die shows odd. What is the probability that I win? Compare this to a fair die (i.e., a die with equal probabilities for each face).

Answer:

$$\begin{aligned}
 P(\text{win}) &= P(1) + P(3) + P(5) \text{ (i.e adding probabilities for all odd numbers)} \\
 &= 0.2 + 0.1 + 0.1 \\
 &= 0.4
 \end{aligned}$$

P(win) of a fair die would be equal, so probability will be for each.

$$\begin{aligned}
 P(\text{win}) \text{ for a fair die} &= P(1) + P(3) + P(5) \\
 &= 3 * (1/6) \\
 &= 0.5
 \end{aligned}$$

ii) What is the entropy of this die? Compare this to a fair die.

Answer:

$$\begin{aligned}
 \text{Entropy} &= - \sum p_i \log(p_i) , \text{ where } i \text{ is from } 1 \text{ to } 6 \\
 &= - [(2 * (0.2) \log(0.2)) + (3 * (0.1) \log(0.1)) + (0.3) \log(0.3)] \\
 &= 0.73645162531
 \end{aligned}$$

$$\begin{aligned}
 \text{Entropy for a fair die is} &= - \sum p_i \log(p_i) , \text{ where } i \text{ is from } 1 \text{ to } 6 \\
 &= - (6 * (1/6) \log(1/6)) \\
 &= 0.77815125038
 \end{aligned}$$