422/622 Midterm 1 Solutions

622 #1

Learning and training are basically the same thing. learning refers to "learning the parameters of a model". So it is the process of building a model or determining the parameters of the model based on the data.

224

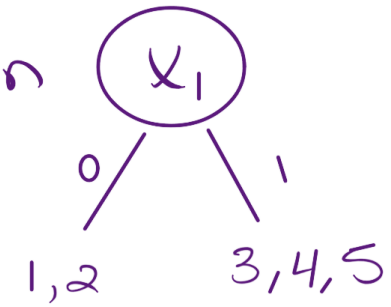
An algorithm is used to train a model (learn the parameters). And a model is used to test new samples using the saved parameters learned as a part of training.

422/622 #2

$$H(S) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}$$

$$= 0.9709$$

Split on x_1



$$H(x_1 = 0) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$$

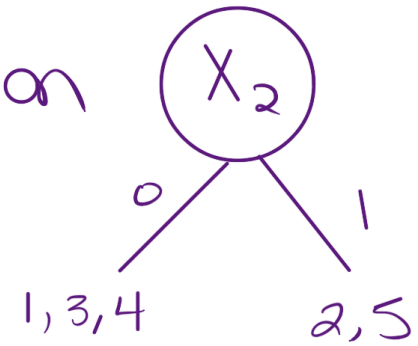
$$H(x_1 = 1) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} =$$

$$= 0.9182$$

$$IG = 0.9709 - \frac{2}{5}(1) - \frac{3}{5}(0.9182)$$

$$IG = 0.019$$

Split on



$$H(X_2 = 0) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}$$

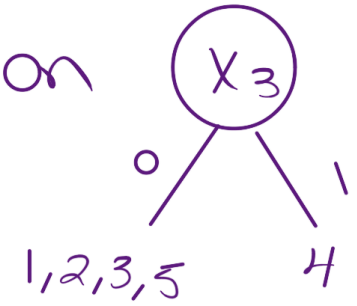
$$= 0.9182$$

$$H(X_2 = 1) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}$$

$$= 1$$

Same IG as X_1

Split on



$$H(X_3 = 0) = \frac{-2}{4} \log \frac{2}{4} - \frac{2}{4} \log \frac{2}{4}$$

$$= 1$$

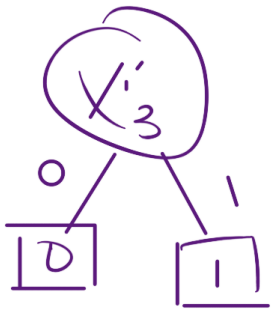
$$H(X_3 = 1) = -\frac{0}{1} \log \frac{0}{1} - \frac{1}{1} \log \frac{1}{1}$$

$$= 0$$

$$IG = 0.9709 - \frac{4}{5}(1) - \frac{1}{5}(0)$$

$$IG = 0.1709$$

So the best feature to split on @ level 1 is X_3



Best depth 1 tree.

422 #3

We can get $\frac{4}{6} = 66\%$ because samples 1 & 2 have the same features and different labels. So we can get only one of those right. That brings us down to $\frac{5}{6}$.

Then we see samples 4 & 5 are conflicting as well so we can only get $\frac{4}{6}$.

622 #3

There are many answers to this question. One would be to perform K-means as normal & use the labels to help you choose K.

So you run K-means w/ the # of classes & then identify "messy" clusters. That is, clusters that have samples from different classes. You can either rerun K-means on the full

data again, increasing k, or you can run it in a hierarchical fashion only on the "messy" clusters. The output could be a set of cluster centers per class.

422 #4 False. Random initialization makes it such that you can get different cluster centers @ the end.

---

622 #4 True. K-means is guaranteed to converge because there are a non-infinite set of clusterings, eventually it will run through them all & stop.

---

622 #5 If k-means uses $k=1$ and k-NN uses $k=N$.

Then they will both output the majority class.

So yes.

Also if $k=N$ for kmeans ? $k=1$ for k-NN you will get the same thing

422 #5 True. When $k = N$ it always outputs the majority class from the training data no matter what the test data is.

---

422 #6 False. If the data is not linearly separable then the perceptron algorithm will loop forever.

---

w22
#6 False. If you run for a fixed # of epochs you can still get a w & b out of the algorithm.

422 #7
622 #7

$$\frac{\partial L}{\partial w} = \frac{1}{\ln 2} \times \frac{1}{1+e^{-y(w \cdot x+b)}} \times -y \times e^{-y(w \cdot x+b)} + \lambda w$$

$$\frac{\partial L}{\partial b} = \frac{1}{\ln 2} \times \frac{1}{1+e^{-y(w \cdot x+b)}} \times -y e^{-y(w \cdot x+b)}$$

$$w = w - \eta \frac{\partial L}{\partial w}$$

$$b = b - \eta \frac{\partial L}{\partial b}$$

422
#8 Depends on where you start.

If you start ≤ $\beta < 0.5$

you will end up @ $\beta = 0$

If you start @ $\beta > 0.5$

you'll end up at ≈ 0.8.

422 #9

622 #8

No. It is underfitting since we're not really clustering at all when $k=1$.

We just average all the points.

So there was something to learn & we didn't learn it.

422 #10

No. A hyperparameter is something the programmer decides.

422 #11
622 #9

$w_1 = 0 \quad w_2 = 0 \quad b = 0$

epoch 1

$a_1 = -3 \times 0 + 0 \times 0 + b = 0 \times y \leq 0$

update

$w_1 = w_1 + xy = 0 - 3 \times -1 = 3$

$w_2 = 0 + 0 = 0$

$b = b + y = -1$

$w_1 = 3 \quad w_2 = 0 \quad b = -1$

$a_2 = 3 \times -2 + 0 \times 0 - 1 = 0 - 7$ ✓

$a_3 = 3 \times -1 + 0 \times -1 - 1 = -4$ ✓

$a_4 = 3 \times 1 + 0 \times 2 - 1 = 2$ ✓

$a_5 = 3 \times 2 + 0 \times 2 - 1 = 4$ ✓