Midterm

Friday, March 25, 2022 8:19 PM

1. Answer the following concept questions (please make your answers concise):
   1) [5 pts] What is the bias-variance trade-off? How to address bias and variance respectively?

The bias-variance trade-off is a common problem in machine learning models. Firstly, bias is the difference between the average prediction of the model & the actual value. Models with high bias lead to underfitting in which the model is not complex enough & end up with high training & testing error. Variance is the variability or the spread of model predictions. Models with high variance lead to overfitting in which the model is too complex & performs very well on the training data but not on the testing data.

Bias & variance are inversely correlated, meaning a model with high variance will have low bias & a model with low variance will have high bias. Hence, there is a tradeoff & a good middle ground must be found. If a model is too simple, it will likely underfit the data resulting in high bias & low variance. However, if a model is too complex, it will likely overfit the data resulting in low bias & high variance.

If a model is too simple & has high bias, the bias can be reduced by adding more parameters, adding more epochs, & overall making the model more complex. On the otherhand, if a model is too complex & has high variance, the variance can be reduced by removing parameters/features, adding regularization, & overall making the model less complex.

   2) [5 pts] What is overfitting? List several common techniques that can reduce overfitting.

Overfitting is a common problem in machine learning where the model fits the training data to closely & fits noise resulting in poor accuracy on new (test) data. This is usually the result of the model being too complex. When a model is overfit it tends to have high variance.

One way to fix this is to make the model simpler by removing parameters/features, adding regularization, or adding more data.

3) [5 pts] What is learning rate? Why learning rate cannot be too large or too small.

Learning rate is a hyperparameter that can be used to tune the step size at each iteration of training. It can be used to speed up or slow down how fast the algorithm moves toward a local/global minimum of a loss function.

If the learning rate is too small, it will take a very long time to reach the optimum. Alternatively, if it is too big, it will take less time to reach the optimum but may skip over it & end up never converging.

2. use the following confusion matrix of a classifier:

		Predicted results: True		Predicted results: False	
Actual values	True	80	TP	30	FN
	False	40	FP	70	TN

1) [5 pts] What is the precision? Please show your calculation.

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{80}{80+40} = \frac{80}{120} = 0.6$$

2) [5 pts] What is the recall? Please show your calculation.

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{80}{80+30} = \frac{80}{110} = 0.72$$

3) [5 pts] What is its F_1-score? Please show your calculation.

$$F_1\text{-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \cdot 0.6 \cdot 0.72}{0.6 + 0.72} = 0.654$$

3. Consider training a decision tree using the training set S as follows:

Sky	Temperature	Wind	EnjoySport
Cloudy	Low	Mild	Yes
Cloudy	High	Strong	No
Sunny	High	Strong	Yes
Sunny	High	Mild	No

yes = p

no = n

1) [4 pts] What is the sample entropy of the training set S?

Entropy → $H = \frac{-p}{p+n}\log_2\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log_2\left(\frac{n}{p+n}\right)$

$= \frac{-2}{2+2}\log_2\left(\frac{2}{2+2}\right) - \frac{2}{2+2}\log_2\left(\frac{2}{2+2}\right)$

$= \frac{-1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$

$= 1$

2) [5 pts] What is the information gain of attribute Sky over S?

$H(S|Sky) = \frac{2}{4}\left(\frac{-1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)$

$+ \frac{2}{4}\left(\frac{-1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1$

$Gain(S|Sky) = 1 - 1 = 0$

3) [5 pts] What is the information gain of attribute Temperature over S?

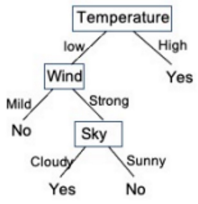
$H(S|Temp) = \frac{1}{4}\left(\frac{-1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right)$

$+ \frac{3}{4}\left(\frac{-1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) = 0.689$

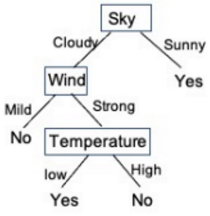
$Gain(S|Temp) = 1 - 0.689 = 0.311$

4) [6 pts] Which of the following is the decision tree we will obtain using the training set S?

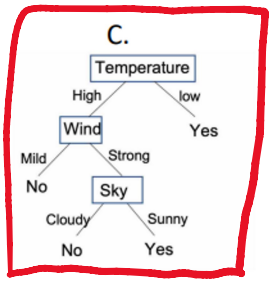
A.



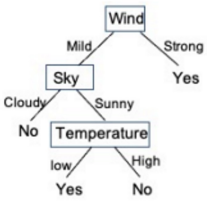
B.



C.



D.



5) [4 pts] What is the predicted result for below example using the decision tree in 4)?
(Sky = Cloudy, Temperature = Low, Wind = Strong)

Based on the decision tree in 4 → EnjoySport = Yes

6) [5 pts] What is the accuracy rate of the decision tree obtained in 4) over training set S?

The accuracy is 100%.

7) [6 pts] Can the tree in 4) be further pruned without losing accuracy? Why?

No it cannot. If you prune the last node 'Sky' and 'Temperature' = High and 'Wind' = Strong, the decision can either be yes or no depending on if 'Sky' is either Cloudy or Sunny.

4. [6 pts] Use the Naïve Bayes Algorithm to predict a new instance based on a dataset with 10 examples below. Show your calculation.

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes

The new instance is: *<Outlook=Rain, Temperature=Mild, Humidity=High, Wind=Weak>*

Naive Bayes' Classifier $\rightarrow v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$

$P(\text{tennis} = \text{yes}) = 6/10$ $\quad P(\text{tennis} = \text{no}) = 4/10$

$P(\text{rain} | \text{tennis} = \text{yes}) = 3/6$ $\quad P(\text{rain} | \text{tennis} = \text{no}) = 1/4$

$P(\text{mild} | \text{tennis} = \text{yes}) = 3/6$ $\quad P(\text{mild} | \text{tennis} = \text{no}) = 1/4$

$P(\text{high} | \text{tennis} = \text{yes}) = 3/6$ $\quad P(\text{high} | \text{tennis} = \text{no}) = 3/4$

$P(\text{weak} | \text{tennis} = \text{yes}) = 5/6$ $\quad P(\text{weak} | \text{tennis} = \text{no}) = 2/4$

$$\Rightarrow P(\text{tennis} = \text{yes} | \text{data}) = P(\text{tennis} = \text{yes}) \cdot P(\text{rain} | \text{tennis} = \text{yes}) \cdot P(\text{mild} | \text{tennis} = \text{yes}) \cdot P(\text{high} | \text{tennis} = \text{yes}) \cdot P(\text{weak} | \text{tennis} = \text{yes})$$

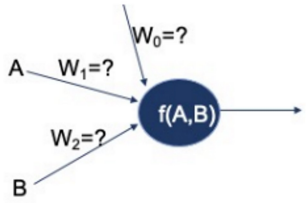
$$= \frac{6}{10} \cdot \frac{3}{6} \cdot \frac{2}{6} \cdot \frac{2}{6} \cdot \frac{5}{6} = 0.0273$$

$$\Rightarrow P(\text{tennis} = \text{no} | \text{data}) = P(\text{tennis} = \text{no}) \cdot P(\text{rain} | \text{tennis} = \text{no}) \cdot P(\text{mild} | \text{tennis} = \text{no})$$

· P(rain | tennis = no)
· P(mild | tennis = no)
· P(high | tennis = no)
· P(weak | tennis = no)
$= 4/10 \cdot 1/4 \cdot 1/4 \cdot 3/4 \cdot 3/4 = 0.0094$

=> Since $0.0273 > 0.0094$ → **Play Tennis = Yes**

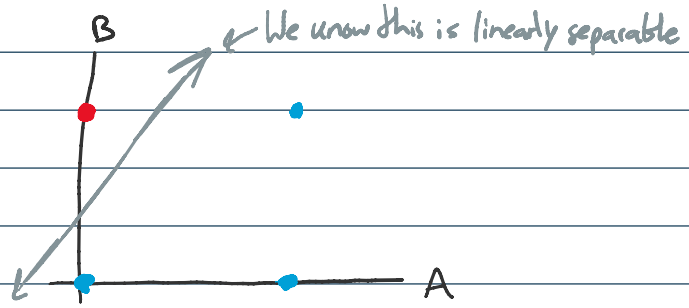
5.Consider the following two-input perceptron. Now we want to choose appropriate weights of the inputs to turn the perceptron into a Boolean function $f(A, B)$.



1) [8 pts] Decide weights (W_0, W_1 and W_2) that make $f(A, B) = A\ OR\ \neg B$. Please show your calculation.

A or ¬B

A	B	¬B	A or ¬B
1	1	0	1
1	0	1	1
0	1	0	0
0	0	1	1



Formula → $W_0 + W_1 \cdot 1 + W_2 \cdot 1 >= 0$ +
$W_0 + W_1 \cdot 1 + W_2 \cdot 0 >= 0$ +
$W_0 + W_1 \cdot 0 + W_2 \cdot 1 <= 0$ −
$W_0 + W_1 \cdot 0 + W_2 \cdot 0 >= 0$ +

Initialize all weights to 1.

For A=1 & B=1 → $1 + 1(1) + 1(0) = 2 > 0$ ✓
For A=1 & B=0 → $1 + 1(1) + 1(1) = 3 > 0$ ✓

For A=1 & B=1 → $1+1(1)+1(0)=2>0$ ✓
For A=1 & B=0 → $1+1(1)+1(1)=3>0$ ✓
For A=0 & B=1 → $1+1(0)+1(0)=1<=0$ ×

$w_0 \rightarrow -1$

For A=0 & B=1 → $-1+1(0)+1(0)=-1<=0$ ✓
For A=0 & B=0 → $-1+1(0)+1(1)=0>=0$ ✓
For A=1 & B=1 → $-1+1(0)+1(0)=-1>=0$ ×

$w_0 \rightarrow 0.5$

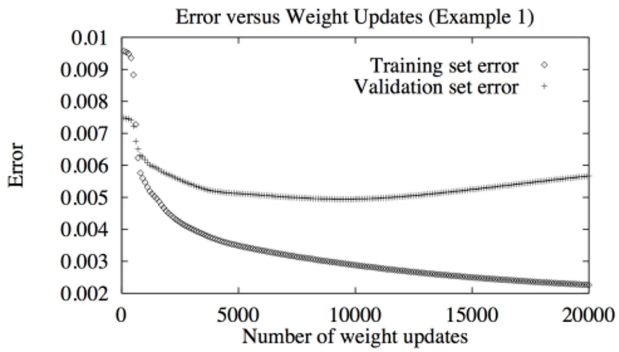
For A=1 & B=1 → $0.5+1(0)+1(0)=0.5>=0$ ✓
For A=1 & B=0 → $0.5+1(1)+1(1)=2.5>=0$ ✓
For A=0 & B=1 → $0.5+1(0)+1(0)=0.5<=0$ ✓

$w_2 = -1$

=> $w_0 = 0.5, w_1 = 1, w_2 = -1$

2) [5 pts] Which of the following the Boolean function *f(A,B)* can NOT represent?
A. AND
B. OR
C. XOR
D. NAND

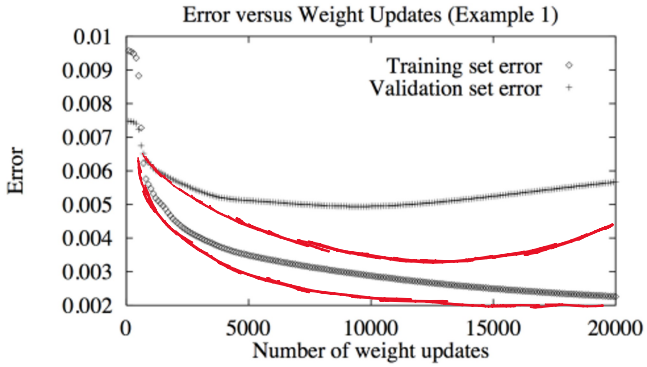
6. Consider below training error and test error observed as we train for a neural network using batch gradient descent.



1) [3 pts] Is there overfitting with the trained model? How do you know?

Yes, there is overfitting in this model. You can tell from the graph. As the number of weight updates increases, the training set error decreases while the validation set error initially decreases & then increases, resulting in a large gap between the errors. One way to prevent this is with early stopping.

2) [3 pts] i) If we double the size of the training data, plot the new curves (on the figure together with the old curves) for training error and testing error, respectively.
[5 pts] ii) Briefly explain why we may have such new curves.



Adding more training data would decrease the variance, resulting in a simpler model that may have more bias but less variance & thus is not as overfitting compared to the previous model.

7. Consider three classifiers that make the following predictions for a sample X. We want to make the final decision using the ensemble of the three classifiers.

Sample x	Result
Classifier 1	Class 2
Classifier 2	Class 2
Classifier 3	Class 1

1) [3 pts] If we use the majority vote as the fusion function, what is our final decision?

Since Class 2 appears twice & Class 1 appears once → Class 2

2) [5 pts] If we use the weighted majority vote as the fusion function, what is our final decision? Given the following weights of the classifiers.
Classifier 1: 0.2
Classifier 2: 0.6
Classifier 3: 0.2

Class 1 → $0.2(0)+0.6(0)+0.2(1) = 0.2$
Class 2 → $0.2(1)+0.6(1)+0.2(0) = 0.8$ => Class 2

3) [7 pts] If we use the Naïve Bayes method as the fusion function, what is our final decision? Given the following confusion matrix of the classifiers. Show you calculation.

i) Classifier 1

	Class1	Class2
Class1	70	10
Class2	30	50

ii) Classifier 2

	Class1	Class2
Class1	80	30
Class2	20	70

iii) Classifier 3

	Class1	Class2
Class1	80	10
Class2	40	40

Bayes → $P(w|x) = \frac{P(x|w)\cdot P(w)}{P(x)}$
$\mu_j(x) \propto \prod_{i=1}^{L} \hat{P}(w_j|d_{i,j}(x)=1)$

Classifier 1: $\hat{P}(w_1|d_{1,2}(x)=1) = 10/60$, $\hat{P}(w_2|d_{1,2}(x)=1) = 50/60$
Classifier 2: $\hat{P}(w_1|d_{2,2}(x)=1) = 30/100$, $\hat{P}(w_2|d_{2,2}(x)=1) = 70/100$

Classifier 1 : $P(w_1|d_{1,2}(x)=1) = \frac{10}{60}$, $P(w_2|d_{1,2}(x)=1) = \frac{50}{60}$

Classifier 2 : $\hat{P}(w_1|d_{2,2}(x)=1) = \frac{30}{100}$, $\hat{P}(w_2|d_{2,2}(x)=1) = \frac{70}{100}$

Classifier 3 : $\hat{P}(w_1|d_{3,1}(x)=1) = \frac{80}{120}$, $\hat{P}(w_2|d_{3,1}(x)=1) = \frac{40}{120}$

$\Rightarrow$ Class1 $= \frac{10}{60} \cdot \frac{30}{100} \cdot \frac{80}{120} = 0.03$

Class2 $= \frac{50}{60} \cdot \frac{70}{100} \cdot \frac{40}{120} = 0.194 \Rightarrow$ Class 2

I pledge my honor that I abided by the Stevens Honor System
-Tim Demetriades