ABE516X - Checking in and reviewing exercise
Answer these questions and turn in on Canvas as a PDF.

Mark Deutsch
ABE 516X
10/24/2019

1. Define supervised and unsupervised learning in machine learning. Give two examples of each.
    a. **Unsupervised learning**: When there is no training set and the goal is to cluster the data and reduce the complexity. In unsupervised learning, there are no classifiers so the goal is not classification of the data. The model can help you identify the trends of the data and uncover how you might go about classifying the data.
        i. Ex. Group machine warranty claims. Give a model a bunch of machine data and then have it group the machines by warranty claims so you can identify the machine data that corresponds to the warranty claim. Then you can pursue different areas to investigate the claims further.
        ii. Ex. Identify high performing machines. Gather sprayer machine data and group them by performance metrics of interest. Then, try to cluster the data and figure out what made the sprayer successful in completing its task with high quality.
    b. **Supervised learning:** When there is a training set and the goal is to classify the data. Supervised learning gives the model a "truth" to "learn" what the different classes are. Then, once the model knows what the different classes are, it can classify data on its own.
        i. Ex. Vision based sense and spray. Train the model on how to classify a weed vs. a plant by giving it a training set that identifies the difference between a plant and a weed. Then, whenever the model sees a weed, spray it.
        ii. Ex. Vision based livestock management. Train a model with pictures of a cow laying down, a cow standing up, or a sick vs. healthy cow. Then, monitor the heath and behavior of your livestock and use it to make decisions.
2. Define bias and variance in terms of how we use it to evaluate a model.
    a. Bias is essentially the metric that shows how well a line can capture the true relationship between two variables. Variance is about how well this line performs with different data sets. We could use these metrics in class to evaluate the quality of the models. A high quality model has low bias and low variability. Overfitting is when you drive the bias really low, but drive the variance very high. The low bias means the line can fit the data set very well, but when another dataset is introduced, it does very bad at fitting. If we see a model that has low bias but high variance, then we know that the model is probably over fit. Ideally, we want both the variance and bias to be low.

3. In terms of bias and variance: The simpler the model, the higher the **__bias__**, and the more complex the model, the higher the **__variance__**.
4. What is the conditional independence assumption in the Naïve Bayes classifier?
    a. This assumption says that the effect of a predictor on a class is independent of other predictors. This means that we can change other predictors all we want, but the effect of the predictor of interest on the class remains the same.
5. What is a confusion matrix and how should it be interpreted?
    a. A confusion matrix lists true positives, true negatives, false positives, and false negatives. This matrix helps you understand how well your machine-learning algorithm is performing. It shows you how many mistakes the algorithm made, by showing you the actual truth and the model prediction. The numbers in this matrix are used directly to calculate performance metrics for the algorithm. As the matrix show below, we would like to minimize the mis-classifications of the data.

		Actual class	
		Cat	Dog
Predicted class	Cat	5	2
	Dog	3	3

In this confusion matrix, of the 8 actual cats, the system predicted that three were dogs,

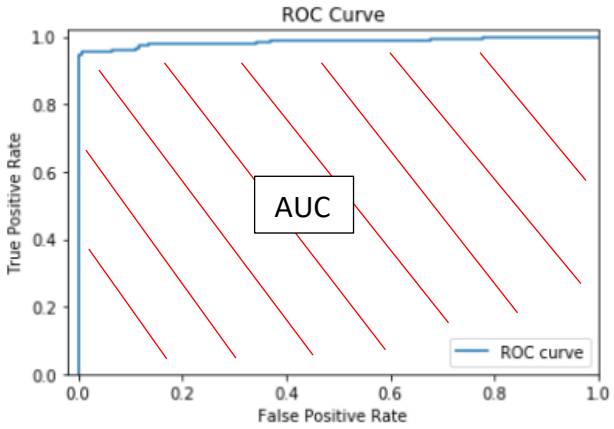
	Class 1 Predicted	*Class 2 Predicted*
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

6. Fill in this table.

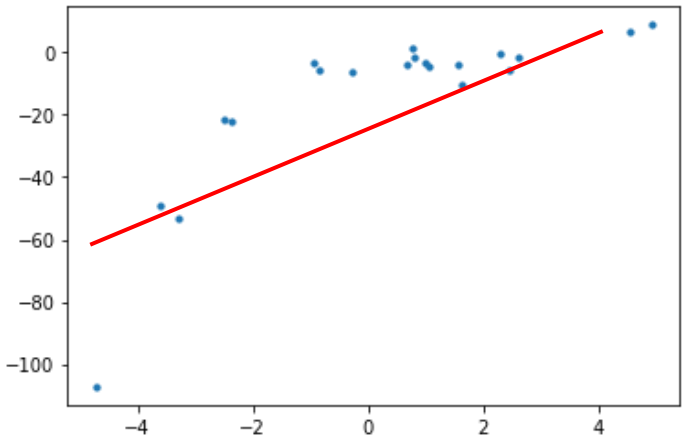
Metric	Formula	Interpretation
Accuracy	(TP + TN)/(TP+TN+FP+FN)	Overall performance of a model
Precision	TP / (TP + FP)	When the model predicts yes, how often is it correct
Recall Sensitivity	TP / (TP + FN)	When the actual is yes, how often does it predict yes
Specificity	TN / (TN + FP)	When the actual is no, how often is it correct
F1 score	2 * precision * recall / (precision + recall)	Combination metric of precision and recall. F1 score is high if there is a balance between precision and recall. It is low if one measure is improved at the expense of the other.

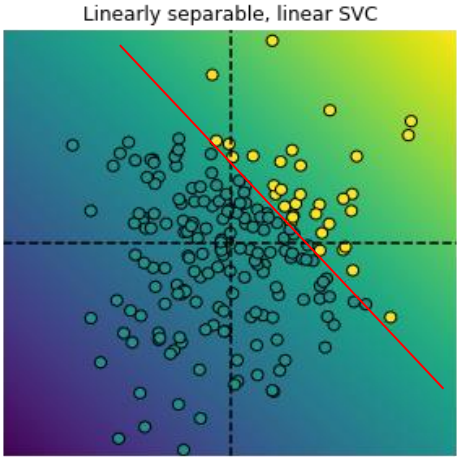
7.

a. An ROC curve is the plot of true positive rate (or recall) vs. false positive rate
b. What is an ideal AUC?
    i. An ideal AUC (area under curve) is 1. This means that the classification algorithm is very good at minimizing false negatives and true negatives. The area is the area under the ROC curve.
c. Draw it on an ROC curve.

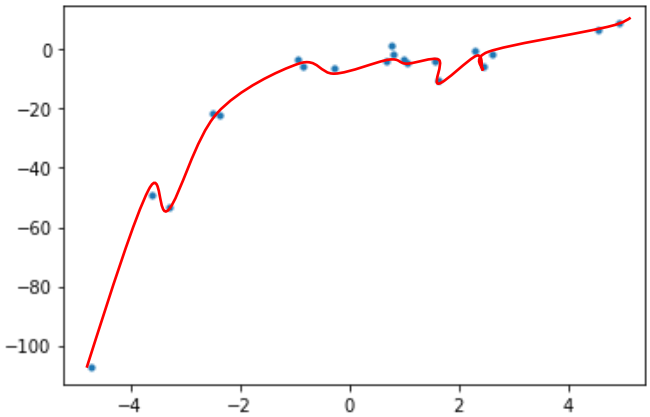


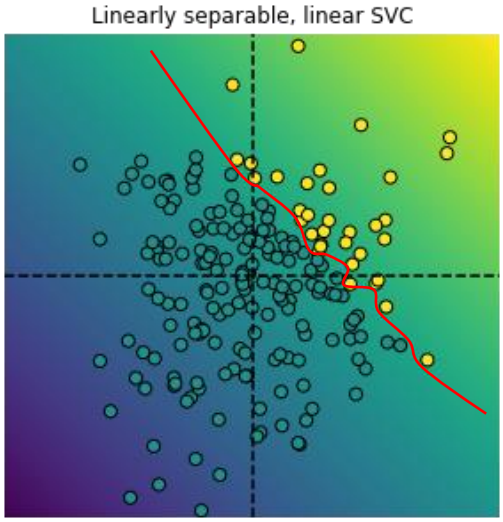
8. Draw an example of underfitting.





Draw an example of overfitting.





9. Identify one thing you've learned in this class that you've found helpful.

- The thing I feel like I have learned most in this class is just data manipulation and data wrangling techniques in general. I have gotten much better at wrangling data and I have applied many of the concepts I have learned in class to my own data sets. In addition, I have learned more about how statistics can really be used in real data and I have found this helpful in my own research. Also, learning the basics of machine learning has helped me become more familiar with what people in industry are doing and I can now understand what machine learning means.