ABE516X - Checking in and reviewing exercise

Answer these questions and turn in on Canvas as a PDF.

You can create space as you need by editing this document.

1. Define supervised and unsupervised learning in machine learning. Give two examples of each.

Supervised learning is a machine learning process under supervision, which the dataset is usually well-labeled. In contrast, the dataset for unsupervised learning is typically not labelled, and allow the machine to self-identify the similarities and differences without the presence of supervisor.

Supervised learning is what we have done so far in the class, the examples include classification (Naïve Bayer, SVM, random forest) and regression. Examples of unsupervised learning are clustering and density estimation.

2. Define bias and variance in terms of how we use it to evaluate a model.

Bias is the difference between predicted and actual values in the training dataset. Variance describes the spread of predicted and actual values in the test dataset.

Over-simplified model often results in high bias but low variance; over-fitted model often results in low bias but high variance.

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?

The NB classifier assumes that the probability of a variable (X) is independent of the other variable (Y).

5. What is a confusion matrix and how should it be interpreted?

Confusion matrix measures the performance of machine learning by comparing the predict and actual outputs. The summary table of confusion matrix include

a) True positive (predicted “yes” when it should be “yes”)
b) False positive (predicted “yes” when it should be “no”)
c) True negative (predicted “no” when it should be “no”)
d) False negative (predicted “no” when it should be “yes”)

6. Fill in this table.

Metric	**Formula**	**Interpretation**
Accuracy	(TP + TN)/(TP+TN+FP+FN)	Overall performance of a model
Precision	TP/(TP+FP)	Proportion of positive identification that was actually correct
Recall Sensitivity	TP/(TP+FN)	Proportion of actual positive that was identified correctly
Specificity	TN/(TN+FP)	Proportion of actual negative that was identified correctly
F1 score	$2 * \frac{precision\ x\ recall}{precision+recall}$	Weighted average of precision and recall

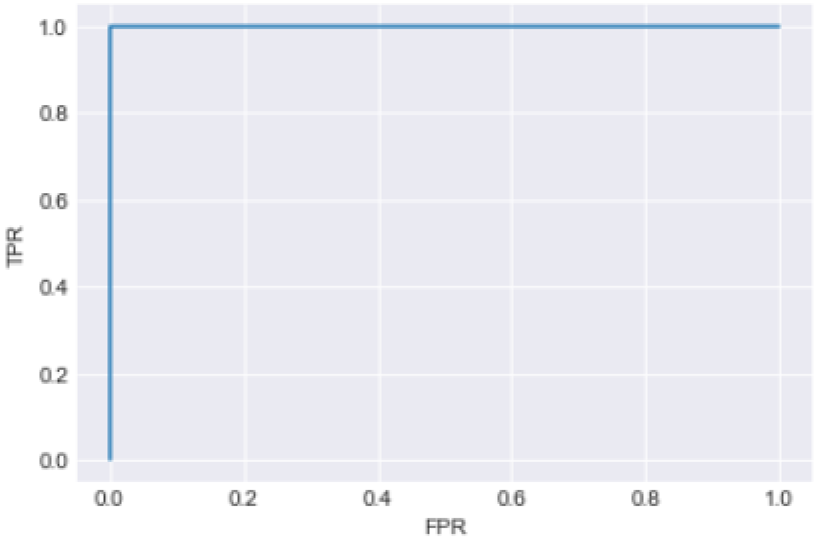
7.

a. An ROC curve is the plot of _true positive rate__ vs. _false positive rate__.

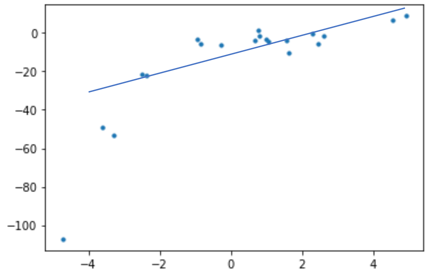
b. What is an ideal AUC?

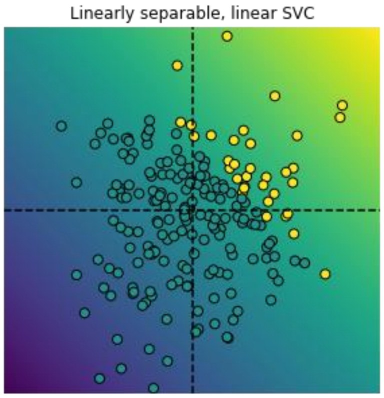
- Steep vertical graph, which the AUC will cover all graph area (1.0)

c. Draw it on an ROC curve.

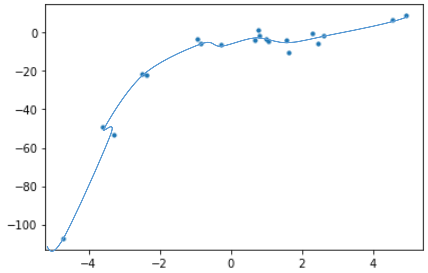


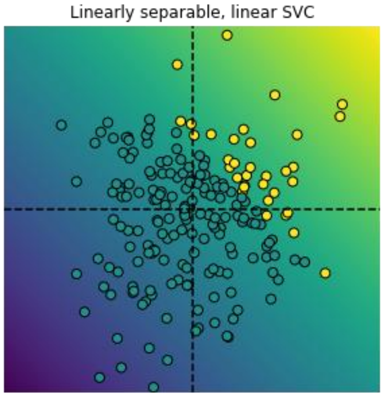
8. Draw an example of underfitting.




Linearly separable, linear SVC

Draw an example of overfitting.




Linearly separable, linear SVC

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

For machine learning: regression and random forest, which could potentially help me to predict missing values due to missing samples. I just need to work on incorporating more labels (i.e. environmental parameters) to improve the prediction.

In general: Python saved me some time in managing and visualizing data. I could pull sub-datasets from the “master dataset” more easily than trying to use “filter” and copy paste function in Excel.