Homework assignments will be done individually: each student must hand in their own answers. Use of partial or entire solutions obtained from others or online is strictly prohibited. Electronic submission on Canvas is mandatory.

**Gaussian Process**: (20 pts) Prove the identity

by multiplying both sides by the matrix

and making use of the definition .

**Graphical Models**: (20 pts)

Consider the table below. Show by direct evaluation that this distribution has the property that a and b are marginally dependent, so that but that they become independent when conditioned on c, so that for both and

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 0 | 0 | 0 | 0.192 |
| 0 | 0 | 1 | 0.144 |
| 0 | 1 | 0 | 0.048 |
| 0 | 1 | 1 | 0.216 |
| 1 | 0 | 0 | 0.192 |
| 1 | 0 | 1 | 0.064 |
| 1 | 1 | 0 | 0.048 |
| 1 | 1 | 1 | 0.096 |

1. Find to show that .
2. Find .
3. Find to show that

**Decision Tree**: (20 pts) Implement a Decision Tree model for the Titanic data set.

1. Explain how you preprocess the features.
2. Build a tree on the training data and evaluate the performance on the test data.
3. Compare Gini index and Information Gain.
4. Report your best accuracy on the test data set.
5. Give a brief description of your observations.

**Boosting**: (20 pts) Implement AdaBoost for the Titanic data set. You can use package/tools to implement your decision tree classifiers. The fit function of DecisionTreeClassifier in sklearn has a parameter: sample weight, which you can use to weight training examples differently during various rounds of AdaBoost.

1. Plot the train and test errors as a function of the number of rounds from 1 through 500.
2. Report your best accuracy on the test data asset.
3. Give a brief description of your observations.

**Random Forest**: (20 pts) Implement Random Forest Classifier for the Titanic data set. You can use package/tools to implement your decision tree classifiers. The fit function of DecisionTreeClassifier in sklearn has an attribute: feature\_importances\_, which you can use to find important features used in Random Forest.

1. Plot the train and test errors as a function of the number of rounds from 1 through 500.
2. Visualize the average of each feature importance and report top 10 important features.
3. Report your best accuracy on the test data asset.
4. Give a brief description of your observations.