15-781 Midterm Example Questions

1 Short Answer

(a) **(True or False?)** If $P(A|B) = P(A)$ then $P(A \wedge B) = P(A)P(B)$.

True

(b) What is the entropy of the following probability distribution: $[0.0625, 0.0625, 0.125, 0.25, 0.5]$?

1 7/8

(c) **(True or False?)** Because decision trees learn to classify discrete-valued outputs instead of real-valued functions it is impossible for them to overfit.

False

(d) **(True or False?)** Assuming a fixed number of attributes, a Gaussian-based Bayes optimal classifier can be learned in time linear in the number of records in the dataset.

True

2 Decision Trees

You are stranded on a deserted island. Mushrooms of various types grow wildly all over the island, but no other food is anywhere to be found. Some of the mushrooms have been determined as poisonous and others as not (determined by your former companions' trial and error). You are the only one remaining on the island. You have the following data to consider.

Example	IsHeavy	IsSmelly	IsSpotted	IsSmooth	IsPoisonous
A	0	0	0	0	0
B	0	0	1	0	0
C	1	1	0	1	0
D	1	0	0	1	1
E	0	1	1	0	1
F	0	0	1	1	1
G	0	0	0	1	1
H	1	1	0	0	1
U	1	1	1	1	?
V	0	1	0	1	?
W	1	1	0	0	?

You know whether or not mushrooms A through H are poisonous, but you do not know about U through W. For the first couple of questions, consider only mushrooms A through H.

(a) What is the entropy of IsPoisonous?

```
-5/8 log 5/8 - 3/8 log 3/8 = 0.9544
```

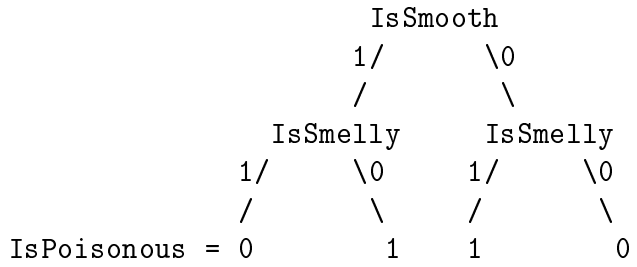
(b) Which attribute should you choose as the root of a decision tree? Hint: You can figure this out by looking at the data without explicitly computing the information gain of all four attributes.

```
IsSmooth
```

(c) What is the information gain of the attribute you chose in the previous question?

```
approximately 0.0487
```

(d) Build a decision tree to classify mushrooms as poisonous or not.



There are other valid solutions since it only asks for ‘‘a’’ decision tree and doesn’t ask for an ID3 decision tree.

(e) Classify mushrooms U, V, and W using this decision tree as poisonous or not poisonous.

U and V both classify as ‘‘not poisonous’’.
W classifies as ‘‘poisonous’’.

Your solution to this might be different depending on the decision tree of the previous question.

(f) If the mushrooms of A through H that you know are not poisonous suddenly became scarce, should you consider trying U, V, and W? Which one(s) and why? Or if none of them, then why not?

The answer we were going for here should have mentioned the small number of examples and that all of U, V, and W can be seen as ‘‘risky’’ due to the small training set. For example, there are other decision trees that are consistent with the training data (other than the one seen in the solution to part d above) for which the classifications of U, V, and W are different.

3 Gaussian Bayes Classifiers

1. Gaussian-based Bayes Classifiers assume that, given n classes, the kth datapoint was generated by first deciding the class of the kth datapoint according to the class prior probabilities, and then choosing the the kth input vector to be generated randomly by a Gaussian distribution with a mean and (usually) a covariance that is dependent on the choice of the class.

   Describe one or more ways in which this assumption could be wrong in practice.