CS 188, Spring 2003, MT 1

1. Mephistopheles, unfortunately, is addicted to gambling, and goes to the local casino every night. Each night he picks a single game to play: either roulette, blackjack, or craps, with respective probability 1/2, 1/3 and 1/6.
Since we all know gambling doesn't pay off, we shouldn't be surprised to learn that Mephistopheles has a net loss on 90% of the nights when he plays roulette, 80% of the nights when he plays blackjack, and 70% of the nights when he plays craps.
You see Mephistopheles on the street one day and he tells you with great excitement that he won at the casino on the previous evening.
a) (15 points) What is your MAP guess as to what game he was playing? (Show your work. Part of the credit will be given for the correct answer, and part for the work shown.)

b) (15 points) What is the probability, given that he won, that he was playing roulette?

2. You have been asked to develop a probability density model of the weights for Pacific ocean mussels based on a collection of mussels collected by local anglers. You have been given the following weight values (in grams) for each of a bunch of mussels: 100 120 121 122.5 99.2 123.1 118.0 99.4 119.7 121.9 121.9 101 103 101 98.

a) (10 points) Your two subordinates, Wild "non-Parametric" Bill and Christophine "Gaussian" LeNerd are arguing that the mussel data should be modeled using a Parzen non-parametric density and a Gaussian density respectively. What type of density would you use and why? (For this question, the right answer is worth half of the credit, and the reason for that answer is the other half. Also, do not use the information in parts b and c of this question to answer part a.)

b) (5 points) Later you learn that this collection of mussels actually represents a mixture of two different species, the Heavy Pacific Mussel and the Light Pacific Mussel. That is, some of the mussels in the collection come from one species, and some from the other. Assuming that no Heavy Pacific Mussel is lighter than 110 grams and that no Light Pacific Mussel is heavier than 110 grams, what is your Maximum Likelihood estimate for the mean weight of the Heavy Pacific Mussel? (Please show work.)

c) (5 points) After having sorted out the mussels into their two species, Wild Bill, Christine LeNerd and Casey Jones each build a probability model of the Light Pacific Mussels only. Wild Bill uses a Parzen estimate of the Light Mussels, using half of the data to build the density, and the other half (the "hold-out set") to compare results for different variance values of the kernels. Christine LeNerd makes a Gaussian probability density by using the maximum likelihood estimate of the mean and variance of the dataand plugging those into the Gaussian probability density formula. Casey Jones makes a Parzen estimate, using all of the Light Mussel data to build a Parzen estimator, and using all of the Light Mussel weight data to find the variance that gives the highest value of the probability density for the weight values. Which one of the three has the worst probability model for the Light Mussel weights, and why?

3. (5 points each) Evaluate the following expressions. If a numerical answer is required and it can be determined, give the numerical answer. If it cannot be determined from the information given, write "cannot be determined".
IMPORTANT: Capital P should be interpreted as "the probability of" and lower case p should be interpreted as "the probability density of". Also, let -A mean the complement of A or "not A".

a)P(A|B)+P(A| B)

b)P(Q| Q)

c)P(C|D)+P(C|D)

d)P(A|B,A)

e) For this problem only (3.e), let A be the outcome of a certain unfair 6-sided die, and B be the outcome of a different unfair 6-sided die. ("Unfair" means that the probability of each roll is not necessarily 1/6.) Evaluate:

$$\sum_{i=1}^{6}\sum_{j=1}^{6} P(A = i)P(B = j|A = i)$$

f) Let a random variable X be distributed according to a Gaussian distribution with mean 0 and variance = 1. Evaluate:

P(X = 2)

4. A propositional logical formula is called satisfiable if there is a model that makes the formula true. (Recall that a model of a formula is an assignment of true or false to each variable so that the formula is true.)
We would like to be able to find models of logical formulae, and one way to find such models is to turn model finding into a search problem. Assuming that there are N boolean variables and that the start state is the model where no variables are assigned values...

a) (3 points) What are the states of our search space?

b) (2 points) What are the operators?

c) (3 points) What are the goal states?

d) (3 points) What is the depth of the search tree (as a function of N)?

e) (2 points) How many states are there in the search space (as a function of N)?

f) (3 points) Will depth-first-search always find a solution if one exists? Why?

g) (3 points) Considering breadth-first-search and depth-first-search, which search strategy would you prefer to use? Why?

5. (1 point) Suppose you have a small amount of "training data" (labeled data) from two different classes and you want to build a classifier. True or false: collecting additional unlabeled data will never help you build a better classifier no matter what sort of classifier you design.