CSE 517 Spring 2012, HW 2 – due date: March 26, in or before class.

Follow the collaboration policy. If not clear, check that out on the course homepage

Two problems below are about when some snow gooses will come out to play on a particular day, depending on temperature, humidity, light and weather (cloud). To simplify our problems, let's just consider the time, which is the target we try to predict, to be in two classes: – (time with – sign, meaning that we ignore the absolute numeric value) and + (otherwise). That is, we want to predict if they gooses will come out early (–) or late (+) based on temperature, humidity, light and weather (cloud). *All the data except the last two items (on 1/23/88 and 1/24/88) are training examples*. We can ignore the two days with missing values (on 1/04/88 and 1/06/88).

DATE	TIME	TEMP	HUM	LIGHT	CLOUD
11/10/87	11	11	78	12.6	100
11/13/87	2	11	88	10.8	80
11/14/87	−2	11	100	9.7	30
11/15/87	−11	20	83	12.2	50
11/17/87	−5	8	100	14.2	0
11/18/87	2	12	90	10.5	90
11/21/87	−6	6	87	12.5	30
11/22/87	22	18	82	12.9	20
11/23/87	22	19	91	12.3	80
11/25/87	21	21	92	9.4	100
11/30/87	8	10	90	11.7	60
12/05/87	25	18	85	11.8	40
12/14/87	9	20	93	11.1	95
12/18/87	7	14	92	8.3	90
12/24/87	8	19	96	12.0	40
12/26/87	18	13	100	11.3	100
12/27/87	−14	3	96	4.8	100
12/28/87	−21	4	86	6.9	100
12/30/87	−26	3	89	7.1	40
12/31/87	−7	15	93	8.1	95
01/02/88	−15	15	43	6.9	100
01/03/88	−6	6	60	7.6	100
01/04/88	−23	5	.	8.8	100
01/05/88	−14	2	92	9.0	60
01/06/88	−6	10	90	.	100
01/07/88	−8	2	96	7.1	100
01/08/88	−19	0	83	3.9	100
01/10/88	−23	−4	88	8.1	20
01/11/88	−11	−2	80	10.3	10
01/12/88	5	5	80	9.0	95
01/14/88	−23	5	61	5.1	95
01/15/88	−7	8	81	7.4	100
01/16/88	9	15	100	7.9	100
01/20/88	−27	5	51	3.8	0
01/21/88	−24	−1	74	6.3	0
01/22/88	−29	−2	69	6.3	0
01/23/88	−19	3	65	7.8	30
01/24/88	−9	6	73	9.5	30

Given the training examples in the snow goose dataset, predict the classes of the two queries (i.e., the data points on 1/23/1988 and 1/24/1988) using

1. (8 points) Bayes optimal classifier? Show your calculation to support your results. (Issue to consider: do we have enough training examples? If not, can we regroup some attribute values?)
2. (7 points) Naïve Bayes classifier? Show your calculation to support your results. (hint: a similar issue as in problem 1 may also need to be considered.)

Given the training examples in the snow goose dataset,

3. (5 points) what is the maximum a posteriori (MAP) hypothesis given the training data?
   (5 points) and how does the data points on 1/23/1988 and 1/24/1988 will be classified under the MAP hypothesis? Show your calculation to support your results. (hint: a similar issue as in problem 1 may also need to be considered.)
4. (5 points) what is the maximum likelihood (ML) hypothesis given the training data?
   (5 points) and how does the data points on 1/23/1988 and 1/24/1988 will be classified under the ML hypothesis? Show your calculation to support your results. (hint: a similar issue as in problem 1 may also need to be considered.)

Problem 5 – (15 points) Consider the bag of candy problem we discussed in class again. The bag in hand might be one of the five types of bags (h_1: 0% L, h_2: 25% L, h_3: 50% L, h_4: 75% L, and h_5: 100% L). The proportions of these five types of bags were 10%, 20%, 40%, 20%, and 10%, respectively. We sampled 10 candies, and they all had L flavor. The figure on the 3rd page of lecture notes showed how our belief of what type of bag that we had changed over the course of our computation; also see lecture notes for the first two steps of the calculation. The key element in our computation was $P(h_i \mid L(j))$, where L(j) represents j L's in a row (e.g., L(3) means LLL). Now assume that after we computed $P(h_i \mid L(k))$, $k < 10$, we had a disk failure and lost the initial h_i, for all i=1, 2, …, 5. (5 points) Describe how we could still proceed to the next step k+1 using what we had already computed. (10 points) Describe the rationale of your method and what formula to use.