Machine Learning: Pattern Recognition
Midterm Exam

Friday September 27, 2013
9:00 - 10:30

Before you start

1. Indicate your name and student number of everything you hand in.
2. On the first page also list the master program you are currently following and your previous education (e.g. bachelor in XX at the U. of YY).
3. You are allowed to use Bishop's book "Pattern Recognition and Machine Learning", printouts of the slides and printouts of the lecture notes as well as cheat sheets. You are allowed to use your iPad but you are not allowed to use the internet.
4. Insight is what we care about: explain your answers! But be concise in your answers: more is not better. Wrong or imprecise elaborations will be counted as mistakes.
5. Good luck!

Questions

1. In a particular city with a population of 500000, it estimated that 500 people have cancer. There is a blood test that 99 times out of 100 correctly diagnoses cancer in patients. It also, unfortunately, misdiagnoses 5% of people that do not have cancer. With this information, answer the following questions:
   (a) What is $p(cancer)$ and $p(notcancer)$? **/3**
   (b) If a patient takes the blood test and it returns positive, what is the probability the patient has cancer? **/3**
2. In the following we assume that we have a conditional model $p(t|x, w)$, with t the target (label), x the input attribute(s) and w the parameters. We also have a prior $p(w)$ over parameters w and a dataset $\{x_i, t_i\}$, $i = 1..N$.
   (a) Define "overfitting" and describe when and why it occurs. **/2**

(b) Explain how you can detect if a model is overfitting and briefly explain ways to avoid it. /2

(c) For this question, write down the precise form of the objective that needs to be maximized. Use the probabilities and dataset above to write out the objective, i.e. your answer should look like $w^* = \arg\max_w O(w, Data)$ where you fill in the details for $O(w, Data)$.

i. Maximum log-likelihood learning /2

ii. Maximum log-likelihood learning with weight decay /2

iii. Maximum posterior learning /2

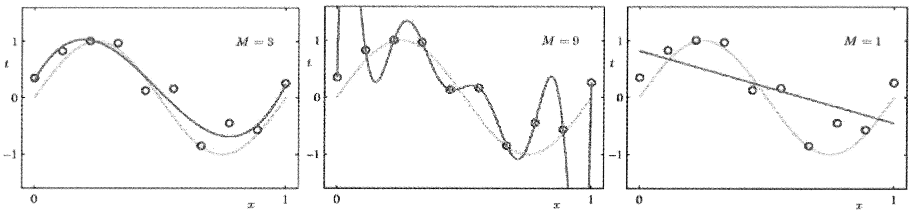
(d) Write down the precise form of the Bayesian predictive distribution $p(t|x, Data)$. Provide as much detail as you can given the definitions for the likelihood, prior and dataset given above. In particular, your answer should not contain any terms (such as the posterior distribution) that remain undefined in terms of the likelihood and prior defined above. This also includes normalization constants. /3

3. Assume a classification problem with two classes C_1 and C_2. Given $p(C_1) = \pi_1$, $p(x|C_1) = \mathcal{N}(u_1, \sigma^2)$, $p(x|C_2) = \mathcal{N}(u_2, \sigma^2)$, answer these questions:

(a) Find a general expression for $p(C_2|x)$. /3

(b) Assuming $\pi_1 = 1/2$, solve for the decision boundary $\hat{x}$, such that $p(C_1|\hat{x}) = p(C_2|\hat{x})$. /3

4. Below are the predictive values (dark solid line) for three model classes (polynomials of degree $M = 3, 9, 1$) trained using a small data set (circles). The predictive values using the true model is light gray.



(a) (b) (c)

(a) Which of the three models has a high bias and which has high variance? /2

(b) The bias-variance is not usually computable for real problems; why is this? /2

(c) Explain how you can (approximately) estimate the bias and variance of a predictive model. /2

(d) If we average the predictions of multiple classifiers, what happens to the variance of our predictions? /2

5. Consider a linear classifier with the following error function with two classes $t = (0, 1)$,

$$E(w, data) = \sum_{i=1}^{N} (t_n - y(x_n, w))^2$$

where $y(x_n, w) = \sum_j w_j x_{jn} + w_0$.

(a) Discuss the disadvantage(s) of this objective for classification problems. **/2**

The logistic regression classifier has a different error function,

$$E(w, data) = -\sum_{i=1}^{N} t_n \log[\sigma(y(x_n, w))] + (1 - t_n) \log[1 - \sigma(y(x_n, w))]$$

where $\sigma(z) = 1/(1 + e^{-z})$.

(b) Does this objective resolve the problems mentioned in the previous question? **/2**

(c) Derive the update equation for *stochastic gradient descent* for logistic regression and describe the algorithm **/4**

(d) Assume you have trained a logistic regression classifier by minimizing the error function above on a training set. Given the attributes of a new test point $\mathbf{x}_*$, explain how you compute the probability that its target value is 1, i.e. $p(t_* = 1|\mathbf{x}_*, w)$ **/3**

(e) Does this probability also reflect our uncertainty in the model parameters? Does it reflect the label noise? **/3**

(f) Is logistic regression a parametric or a non-parametric method? **/2**