

NICTA

## COMP4670/6467

### Introduction to Statistical Machine Learning

#### Assignment 1

Maximum marks	50
Weight	50% of the total assignment marks
Submission deadline	Wednesday, 2 May 2007, 13:00
Submission mode	On paper to M. Hutter or S. Günter or electronic
Estimated time	2-3 hours per lecture week $\approx$ 20min per mark
Late Penalty	20% per day
Some solutions	30.May'07 13:00-16:00

Qinfeng (Javen) Shi will present solutions to some of the exercises on 30.May'07 13:00-16:00 in the Geology Theatre.

## SML Overview & How to Predict (by Marcus Hutter)

**MH1 (1/50)** An arts, a maths, a statistics, a physics and a philosophy student are asked the same question "What is the probability that the sun will rise tomorrow"? How do you think each person answered and why?

**MH2 (1/50)** What is the next number in the following sequences?

(a) 1, 2, ...

(b) 1, 2, 3, 4, ...

(c) 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, ...

(d) 3, 1, 4, 1, 5, 9, 2, 6, 5, 3, ...

Can these have other continuations, other than the obvious ones? How many numbers are enough to be sure of a continuation?

**MH3 (3/50)** Prove Bayes' rule using (only) the axioms of probability.

*Probability axioms:*

(i)  $P(\emptyset) = 0 \leq P(A) \leq 1 = P(\Omega)$ .

(ii)  $P(A \cup B) + P(A \cap B) = P(A) + P(B)$ .

(iii)  $P(A|B)P(B) = P(A \cap B)$ .

*Bayes' rule:* Let  $D$  be a possible event ( $P(D) > 0$ ) and  $H_i$  be a set of mutually exclusive hypotheses ( $H_i \cap H_j = \emptyset \forall i \neq j$  and  $\cup_{i \in I} H_i = \Omega$ ).  $P(H_i)$  is a priori plausibility of hypothesis  $H_i$ ,  $P(D|H_i)$  is the likelihood of event  $D$  under hypothesis  $H_i$ . Then the posterior plausibility of hypothesis  $H_i$  is  $P(H_i|D) = \frac{P(D|H_i)P(H_i)}{\sum_{i \in I} P(D|H_i)P(H_i)}$ .

**MH4 (2/50)** Assume the prevalence of a certain disease in the general population is 1%. Assume there exists a quite reliable test for the disease, say, the test on a diseased/healthy person is positive/negative with 99% probability. If the test (on some randomly selected person) is positive, what is the chance that (s)he has the disease? Explain the result. Hint 1: Use Bayes' rule. Hint 2: the chance is not high!

**MH5 (2/50)** In the game of Chuck-a-Luck three dice are thrown and a player bets some amount on a number (between 1 and 6). He is then rewarded with \$1 for every time his number appears (i.e. 0\$-3\$). What is the proper amount the player should pay in order that this be a fair game?

**MH6 (2/50)** Envelope Paradox:

I offer you two closed envelopes, one of them contains twice the amount of money than the other. You are allowed to pick one and open it. Now you have two options. Keep the money or decide for the other envelope (which could double or half your gain).

Symmetry argument: It doesn't matter whether you switch, the expected gain is the same.

Refutation: With probability  $p = 1/2$ , the other envelope contains twice/half the

amount, i.e. if you switch your expected gain increases by a factor  $1.25 = (1/2) \times 2 + (1/2) \times (1/2)$ .

Present a Bayesian solution.

**MH7 (2/50)** Confirmation Paradox:

- (i)  $R \rightarrow B$  is confirmed by an  $R$ -instance with property  $B$
- (ii)  $\neg B \rightarrow \neg R$  is confirmed by a  $\neg B$ -instance with property  $\neg R$ .
- (iii) Since  $R \rightarrow B$  and  $\neg B \rightarrow \neg R$  are logically equivalent,  $R \rightarrow B$  is also confirmed by a  $\neg B$ -instance with property  $\neg R$ .

Example: Hypothesis ( $o$ ): All ravens are black ( $R$ =Raven,  $B$ =Black).

- (i) observing a Black Raven confirms Hypothesis ( $o$ ).
- (iii) observing a White Sock also confirms that all Ravens are Black, since a White Sock is a non-Raven which is non-Black.

This conclusion sounds absurd.

Present a Bayesian solution.

## Regression and Classification (by Simon Günter)

### SG 1 (8/50)

In the lecture classification using logistic regression was presented.

- $P(y = y_i|x) = \frac{e^{f_i(x)}}{\sum_j e^{f_j(x)}}$
- $L = -\sum_{i=1}^m \log(P(y_i|x_i)) = -\sum_{i=1}^m \log\left(\frac{e^{f_{y_i}(x_i)}}{\sum_j e^{f_j(x_i)}}\right)$

The task is now to implement a classifier using logistic regression

- $f_j(x) = w_j^T x = w_{j1}x_1 + \dots + w_{jn}x_n$
- $\frac{\partial}{\partial w_j} L = -\sum_{i=1}^m (c(y_i == j) - P(y_j|x_i))x_i$   
with  $c()$  indicator function
- Use gradient descent:  $w_i = w_i - \eta \cdot \frac{\partial}{\partial w_i} L$
- First coefficient of  $x$  is always 1 (to code the intercept), i.e. the dimension of the input space increases by 1

The classifier should be tested on the Fisher Iris data set (available on the course home page<sup>1</sup>)

- Sequence of patterns; 4 features and 1 class label  $\in \{1, 2, 3\}$  ( $\rightarrow$  5 dimensional input space)
- Please note that the optimal gain rate  $\eta$  may be very small (much smaller than  $10^{-4}$ !)

The solution of the assignment should include

- Listing of the program (all not exotic prog. lang allowed)
- Output including value of L after each iteration and the final error rate on the whole set

---

<sup>1</sup><http://sml.nicta.com.au/Education/Teaching/IntroToSML/view>

## Unsupervised Learning (by Simon Günter)

### SG 2 (7/50)

*K*-means clustering

- Implement *k*-means and provide a listing of your code
- Apply *k*-means to Fisher Iris data set. Vary *k* from 1 to 10 and repeat the algorithm 100 times for each *k*
- Report for each *k* the lowest found mean square error (MSE)
- Report and comment on anomalies of the evolution of MSE when increasing *k*

## Validation (by Simon Günter)

### SG 3 (7/50)

- Implement *k*-NN and CV for Fisher Iris data set (in case of a tie, simply pick the first class)
- Apply 2-fold, 5-fold and 10-fold CV to select best *k*
- In case of several *k* having lowest error rate, we pick the largest one. Explain why this is a good strategy.
- Report the results for  $k = 1, 3, \dots, 37, 39$
- The optimal errors decrease with the fold number. Explain why.
- The optimal *k* increases with the fold number. Explain why.

## Density Estimation (by Nic Schraudolph)

In this section we'll try various density estimation techniques on the Fischer Iris data set (from the course website). This is unsupervised learning, so make sure you don't use the class labels (5th column) until you get to NS5! :-)

You may use a computing environment that provides linear algebra, but obviously should implement the density estimation algorithms yourself and not just call some library's `FitGaussian()` method. . .

### NS1 (1/50): Parametric

Calculate and report the mean and covariance matrix for the 4-D Gaussian that best fits the data in the maximum likelihood sense. Calculate and report the log-likelihood of the data under this model.

### NS2 (3/50): Nonparametric

Now model the data density with Parzen windows, using isotropic (*i.e.*, spherical) Gaussian kernels of fixed standard deviation  $\sigma = 3$ . Provide a listing of your code, and report the resulting log-likelihood of the data.

### NS3 (8/50): Semiparametric

Now fit a mixture of 3 multivariate Gaussians to the data, using the EM algorithm to fit their means and covariance matrices. Provide a listing of your code. Plot the log-likelihood of the data against the EM iteration number, for 5 restarts from different random initial conditions (*i.e.*, means). Do you find the same solution each time?

### NS4 (1/50): Comparison

Compare the above 3 methodologies in terms of expressive power (*i.e.*, achieved log-likelihood), ease of implementation, and computational cost (consider how that scales to large data sets). When would you use which?

### NS5 (2/50): Classification from Mixture Density

For one of your EM solutions, assign each data point to its most likely mixture component (Gaussian). Now find the mapping (not necessarily one-to-one!) from mixture components to class labels (5th column of the data set) that, when used as a classifier, minimizes the misclassification rate on the data set. Report the mapping, describe the algorithm you used to find it, and the resulting misclassification rate. What do you think of the classification performance?