

COMP4670/6467

Introduction to Statistical Machine Learning

Assignment 1

Maximum marks	15
Weight	15% of final grade
Submission deadline	Wednesday, April 30, 2008, 23:59 (no exceptions)
Submission mode	Email to Scott.Sanner (@nicta.com.au)

Solutions presented in tutorial.

Background

1 (1/15) 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 $P(H_i|D) = \frac{P(D|H_i)P(H_i)}{\sum_{i \in I} P(D|H_i)P(H_i)}$.

2 (.5/15) 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!

3 (1.5/15)

- a) We believe samples $D = \{4, 7, 2, 8\}$ were generated from the uniform distribution:

$$p(x|\theta) \sim U(0, \theta) = \begin{cases} 1/\theta & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

What is the maximum likelihood estimate of θ , i.e., $\arg \max_{\theta} p(D|\theta)$?

- b) Assume the data samples D are received one-by-one as x_i for $i = 1 \dots 4$ and let $D^i = \{x_1, \dots, x_i\}$. Derive the posterior $p(\theta|D^n)$ from prior $p(\theta|D^{n-1})$

$$\begin{aligned} p(\theta|D^n) &= p(\theta|x_n, D^{n-1}) = \frac{p(x_n|\theta, D^{n-1})p(\theta|D^{n-1})}{\int_{\theta} p(x_n|\theta, D^{n-1})p(\theta|D^{n-1})d\theta} \\ &\propto p(x_n|\theta)p(\theta|D^{n-1}) \end{aligned}$$

where we assume $p(\theta|D^0) = U(0, 10)$. What is the posterior distribution of θ given D^4 and $p(\theta|D^0)$, i.e., $p(\theta|D^4)$?

- c) What is the *maximum a posteriori* (MAP) estimate of θ given D^4 and $p(\theta|D^0)$, i.e., $\arg \max_{\theta} p(\theta|D^4)$?
- d) How do the maximum likelihood estimate of θ and the posterior distribution over θ compare in the infinite limit of data? Consider the case where the prior was non-zero for the true value of the parameter θ^* and the case where it was not (i.e., the prior was incorrect).

Regression and Classification

4 (4/15)

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

- 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 η may be very small (much smaller than 10^{-4} !)

The solution of the assignment should include

- Listing of the program (if not using Matlab/Octave, please talk to instructor)
- Output including value of L after each iteration and the final error rate on the whole set

Unsupervised Learning

5 (4/15)

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

6 (4/15)

- 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.

Neural nets

7 (4 pts extra credit on this assignment only, up to max grade 15/15)

See the following web page for a neural network approach to face detection:

<http://users.rsise.anu.edu.au/~ssanner/Software/Vision/Project.html>

The code referenced on this page requires the Matlab neural net toolbox. Since the Mathworks charges money for its neural net toolbox, you get to code up your own version of a multilayer neural net trained by backpropagation for use with this face detection code!

There are three functions that you will have to implement:

- `newff.m`: Creates a new feed-forward neural network. You can ignore some input parameters and assume a 3-layer neural net with one output unit (hidden and output units are sigmoidal), but you must allow the number of input and hidden units to vary as specified by the parameters.
- `train.m`: Trains a neural net for one iteration by backpropagation gradient descent given data samples as column vectors and output target values in corresponding columns.
- `simnn.m`: Given a setting of inputs, feeds all values forward to determine the output of the network given the current weight settings.

The solution should include:

- Listing of commented code for the above three functions.
- A plot of training and validation error on each iteration of training.
- Pictures showing example detection of faces with a trained neural net on a test image for different threshold values. Include the relevant training and testing parameters used to generate these pictures.
- What do you think is the most important way in which the face detector could be improved? Briefly justify your answer.