Introduction to Machine Learning - Homework 1

Prof. Alan Yuille

Spring 2014

Due on Thursday 24/April. 2014. Hand in hardcopy in class.

Question 1. Bayes rule.

A prize is hidden behind one of three doors A,B, and C. The contestant picks a door, say A, but it is left closed. The host opens door C and shows that there is no prize behind it. Should the contestant change his mind and select door B? Formulate this problem as Bayes inference. What should the contestant's decision rule be?

Question 2. Bayesian decision theory

Describe the Bayes risk for making a binary decision $y \in \{-1, +1\}$. How does it depend on the prior and the loss function?

Suppose somebody tosses a fair (unbiased) coin and if the result is 'heads' you get nothing, otherwise you get 5 dollars. How much would you pay to play this game? What if you would win 500 dollars instead of 5 dollars? (Note, there is no "correct" answer to this question).

Suppose you have vector-valued data $\{\vec{x}_i : i = 1, ..., N_1\}$ and $\{\vec{x}_i : i = 1, ..., N_{-1}\}$ from two classes $y \in \{+1, -1\}$. Describe how to learn Gaussian distributions for the distributions $P(\vec{x}|y = +1)$ and $P(\vec{x}|y = -1)$. What is the log-likelihood rule for classifying the data? Hence compute the decision rule. When is this decision rule

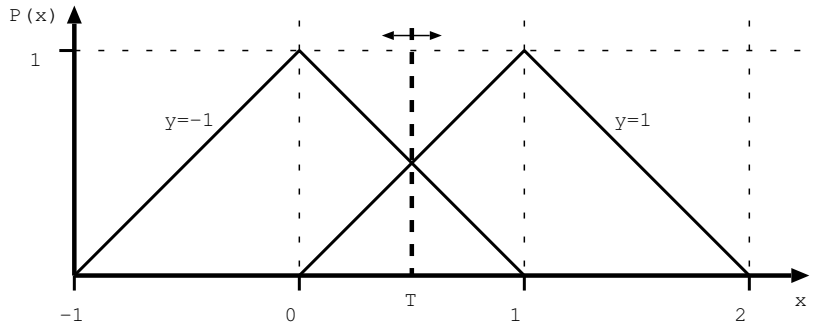
the same as separating the data by a plane?

Question 3. ROC curves and PR curves

Describe the Receiver Operator Characteristic (ROC) curve and the Precision Recall (PR) curve. If you know the ROC curve can you deduce the PR curve? And vice versa? If not, what extra information do you need?

Discuss briefly if you would prefer a ROC or a PR curve for a problem where the number of distractors is much bigger than the number of targets.

Why does Bayes decision theory (for binary classification) reduce to a decision rule of form $\hat{\alpha}_T(x) = -1$, if $x < T$, otherwise $\hat{\alpha}(x) = 1$? How does the threshold T depend on the prior and the loss function? For the likelihood functions shown in the figure, and a threshold $T = 0.5$, calculate the true positive and the false positive rates. (Hint: this reduces to calculating the area of triangles).



Change T to $-1, -0.5, 0, 1, 1.5, 2$ and calculate the rates for all those thresholds. Then, plot the ROC curve. Suppose you have a dataset with 10 samples from $p(x|y = 1)$ and 100 samples from $p(x|y = -1)$. Plot the Precision-Recall curve.

Question 4. Bias and Variance

Let X be a random variable representing the current time (not distinguishing between a.m. and p.m.). Suppose X is sampled from a uniform distribution $X \sim U[0, 12]$. Consider two estimators of the current time $x \in X$:

- $\hat{x}_1$ is a stopped clock (which always shows the same time). Note that it gives the correct estimate twice a day.

- $\hat{x}_2$ is a clock which works with perfect precision but is in the wrong timezone, so it is always one hour late and never gives the correct estimate.

Calculate the bias and the variance of both estimators $\hat{x}_1$ and $\hat{x}_2$. Hint: it may help to formulate this problem in terms of the difference between the current time and the estimate: so that $X \in (-6, 6)$ and the current time is always 0.

Question 5. Memorization and Generalization

Define the risk and the empirical risk. What is the empirical risk introduced? What is the relationship between the risk and the empirical risk as the number of samples tends to infinity?

What is memorization and generalization? How to test generalization?

Question 6. Maximum likelihood and exponential family

Consider an exponential distribution $P(x|\lambda) = \frac{1}{Z[\lambda]} \exp\{\lambda \cdot \phi(x)\}$. Suppose you have data $\mathcal{X} = \{x_1, x_2, ..., x_N\}$, how do you estimate the parameters λ of the distribution using maximum likelihood?

The Bernoulli distribution has form $P(x) = \theta^x(1-\theta)^{1-x}$, where $x \in \{0, 1\}$ (i.e. x takes value 0 or 1) and $\theta \in [0, 1]$. Re-express the Bernoulli distribution as an exponential distribution. And show how to estimate its parameter using maximum likelihood.

Question 7. Curse of Dimensionality

Suppose we want to estimate a probability distribution $f(x)$ in the unit hypercube $\mathbf{R}^d$ from n data samples. If $f(x)$ is complicated, we need many samples to learn it well.

(a) Let n_1 denote the number of samples required to estimate $f(x)$ in $\mathbf{R}^1$. How

many samples are needed for the same density in $\mathbf{R}^d$? If $n_1 = 100, d = 20$, what sample size is needed in this high-dimensional space? (Note, assume that you do not know a parametric form for $f(x)$).

(b) Write a formula for $l_d(p)$, the length of the edge of the hypercube in d dimensions that contains a fraction p of points ($0 \leq p \leq 1$) in the unit hypercube. To better understand the implications of your result, calculate: $l_5(0.01), l_5(0.1), l_{20}(0.01), l_{20}(0.1)$.