

Voor Anne

Eriz Tjerk

Machine Learning: Pattern Recognition Mock Exam Questions

December 8, 2008

The exam will consist of four main questions. One question will be on general knowledge, and the three other will be on a specific method or technique we've seen in the course. Below are a few example questions such as they might be asked on the exam.

You will be allowed to bring and use Bishop's book.

1 General knowledge

Answer the following questions in a few sentences:

1. Explain the terms "training set", "validation set", "test set" and briefly describe what they are used for.
2. What are basis functions, what is the purpose of using them and how can we learn basis functions from data?
3. Let the probability of N independent and identically distributed binary samples $\{x_1, \dots, x_N\}$, $x_n \in \{0, 1\}$ be given by the Bernoulli distribution

$$p(x_i) = \mu^{x_i}(1 - \mu)^{(1-x_i)} \quad (1)$$

Prove that the maximum likelihood estimator of μ is given by

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^N x_n \quad (2)$$

Hint: find the maximum of the log-likelihood rather than the likelihood, since the log is a monotonically increasing function the argument for which the log-likelihood is maximal is the same as for which the likelihood is maximal.

2 k Nearest Neighbours

1. Draw the discriminant k-nearest-neighbours would create with the training data depicted in figure 1, for $k = 1$.
2. The crosses represent class C_1 and the circles class C_2 . If we now get a new sample at $(2.5, 0.5)$, how would it be classified?
3. What is the point of using kNN with $k > 1$? What happens as k becomes bigger. Why not choose k as big as possible?
4. k-Nearest-neighbours gives an estimate of the data density around the new data point. Explain how this is done and what the (contradictory) conditions are that must be satisfied for that estimate to be accurate?

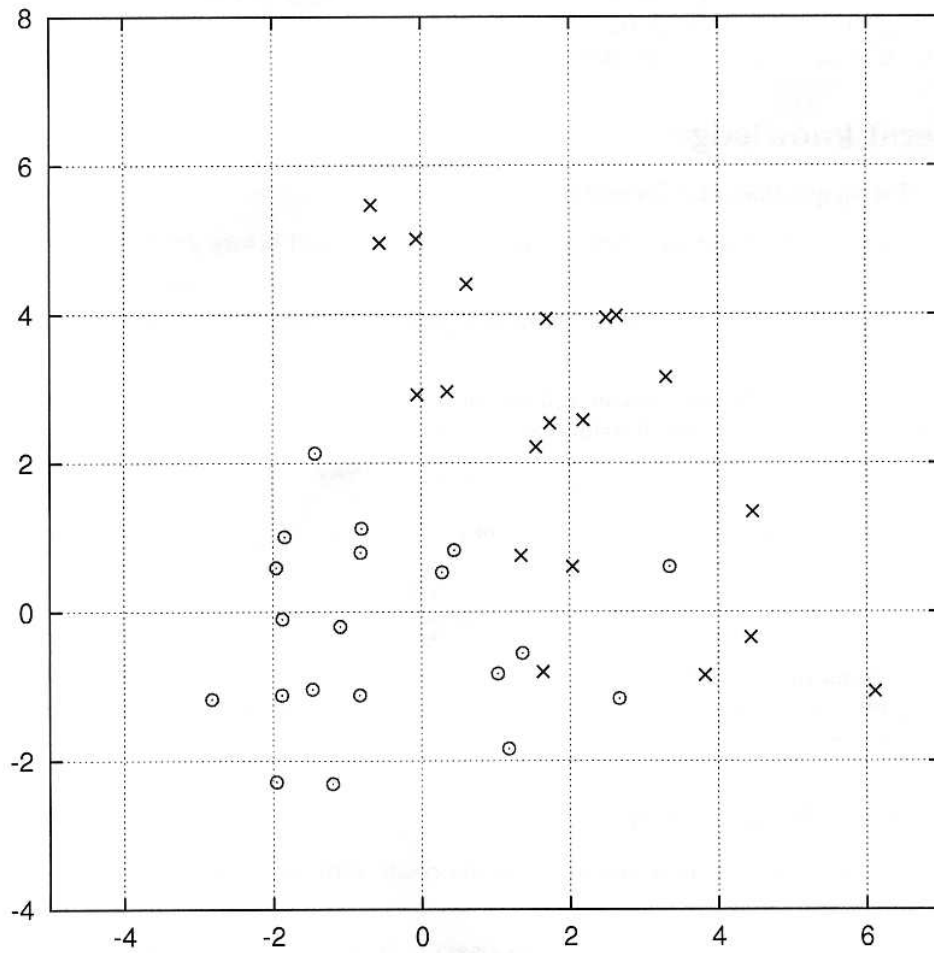


Figure 1: Data for k-nearest-neighbours

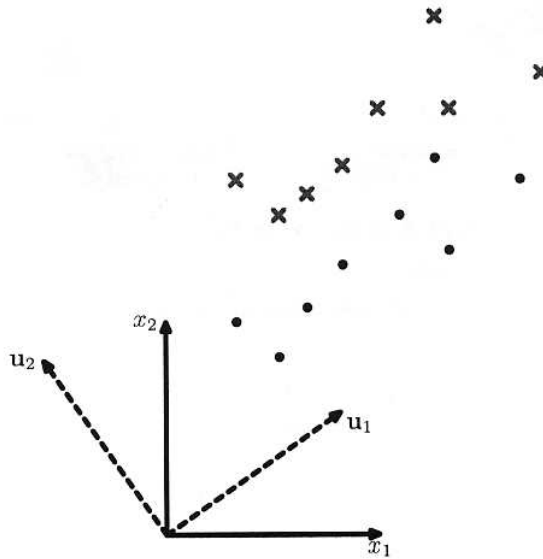


Figure 2: Data for the PCA exercise

3 Principal Component Analysis

PCA is used to project high-dimensional data to a single dimension.

1. Suppose we have a projection vector \mathbf{u} of unit norm. The projection of a datapoint \mathbf{x} on \mathbf{u} is given by the scalar p . Express p in function of \mathbf{u} and \mathbf{x} .
2. The data set is given in the form of a matrix \mathbf{X} , where element x_{rc} denotes the c th dimension of the r th datapoint. That is, each row i of \mathbf{X} is a vector of a datapoint \mathbf{x}_i^T . Suppose that the data has zero mean.
 - (a) Write the projected data — that is, the vector \mathbf{p} — in function of \mathbf{X} and \mathbf{u}
 - (b) Write the variance in p in terms of \mathbf{X} and \mathbf{u}
3. Explain what PCA does if high-dimensional data is projected to a lower space:
 - (a) What is optimised?
 - (b) How do we find the best projection directions? Explain in words.
 - (c) Show mathematically that these are the best projection vectors.
4. What is an auto-encoder? What is the relationship between PCA and auto-encoders?
5. PCA can be described as a probabilistic model, and is then called probabilistic PCA.
 - (a) Give the model of PPCA
 - (b) What are the advantages of considering PCA as a probabilistic model?
6. One popular application PCA is to obtain low-dimensional representations of faces, called eigenfaces.
 - (a) Describe how the data is represented before PCA can be performed
 - (b) What difficulty arises when performing PCA on such very high-dimensional data?
 - (c) Show mathematically how you can obtain the principal components of the face images without computing the covariance matrix explicitly.

7. If we have two-dimensional data as depicted in figure 2 which we want to project down to 1 dimension using PCA.

- (a) What projection vector will PCA choose, \mathbf{v}_1 or \mathbf{v}_2 ?
- (b) Is that optimal for classification?
- (c) How would Fisher's linear discriminant choose the projection vector?