sf2935 Modern Methods of Statistical Learning: 35 Questions to be Considered for the Exam on Thursday the 12th of January, 2017, 08.00 - 13.00

Presented on 13/12/2016

Timo Koski, Pierre Nyquist, Jimmy Olsson & Tetyana Pavlenko,

1 Introduction

This document contains a set of questions/problems on the topics treated in sf2935 Modern Methods of Statistical Learning during the period 2 of 2017. Five of these will be selected to constitute the written exam on Thursday the 12th of January, 2017, 08.00 - 13.00.

The answers/solutions can be produced by study of the relevant chapters in the course textbook *An introduction to Statistical Learning*, by G. James, D. Witten, T. Hastie, R. Tibshirani. Springer Verlag, 2013, and by similar study of the lecture slides on the webpage
`https://www.math.kth.se/matstat/gru/sf2935/statlearnmaterial2016`
In addition, some proficiency in manipulating basic calculus, probability, linear algebra and matrix calculus is required.

This same set of questions (maybe some will be removed and new added) will be valid in the re-exam. Hence we shall NOT provide a solutions manual.

2 The Assignments

1. Suppose that we have a training set consisting of a set of points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and real values y_i associated with each point x_i. We assume that

   $$y_i = f(\mathbf{x}_i) + \epsilon,$$

   where the noise, ϵ, has zero mean and variance σ^2 .

   We want to find a function $\hat{f}(\mathbf{x})$, that approximates the true function $y = f(\mathbf{x})$. We make "as well as possible" precise by measuring the mean squared error between y and $\hat{f}(\mathbf{x})$ which we want to be minimal both for $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

   We can decompose its expected error on an unseen sample $\mathbf{x}$ as follows:

   $$\mathrm{E}\Big[\big(y - \hat{f}(\mathbf{x})\big)^2\Big] = \mathrm{Bias}\big[\hat{f}(\mathbf{x})\big]^2 + \mathrm{Var}\big[\hat{f}(\mathbf{x})\big] + \sigma^2 \tag{2.1}$$

   $$\tag{2.2}$$

   Where:

   $$\mathrm{Bias}\big[\hat{f}(\mathbf{x})\big] = \mathrm{E}\big[\hat{f}(\mathbf{x})\big] - f(\mathbf{x}) \tag{2.3}$$

and

$$\mathrm{Var}\big[\hat{f}(\mathbf{x})\big] = \mathrm{E}\Big[\big(\hat{f}(\mathbf{x}) - \mathrm{E}[\hat{f}(\mathbf{x})]\big)^2\Big] \tag{2.4}$$

a) Deduce these equations.

b) Write a small essay on the *Bias -Variance Trade-Off* in terms of these equations.

2. Rosenblatt's perceptron algorithm is written in a pseudocode as

1: $\mathbf{w}_0 \leftarrow \mathbf{0}$, $b_0 \leftarrow 0$, $k \leftarrow 0$ $\mathcal{R} \leftarrow \max_{1 \le i \le l} \|\mathbf{x}_i\|$
2: **repeat**
3: **for** $i = 1$ to l **do**
4: **if** $y_i \left(\ll \mathbf{w}, \mathbf{x}_i \gg + b_k\right) \le 0$ **then**
5: $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + y_i \mathbf{x}_i$,
$b_{k+1} \leftarrow b_k + y_i \mathcal{R}^2$,
$k \leftarrow k + 1$
6: **end if**
7: **end for**
8: **until** no mistakes made in the loop
9: **return** $\mathbf{w}_k, b_k$, where k is the number of mistakes.

a) Explain in plain words what the algorithm does.

b) Formulate the Perceptron Convergence Theorem (Novikoff 1962) and interpret it in terms of the perceptron algorithm and its geometry.

3. Explain the concept of separating hyperplane classification. Specifically, consider the case where the feature space $\mathcal{X}$ is p-dimensional and the output space is $\mathcal{Y} = \{-1, 1\}$. Describe how a separating hyperplane is defined and how such a (generic) hyperplane can be used for classification. Define the maximal margin hyperplane and describe in some detail how it can be found for a given training set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$.

4. Consider a p-dimensional feature space $\mathcal{X}$, output space $\mathcal{Y} = \{-1, 1\}$ and a training set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Define the support vector classifier; you must explain all the quantities involved. What is the

difference between the support vector classifier, i.e. the associated hyperplane, and the maximal margin classifier? What does the term "support vector" refer to and why is this an important concept for this particular classifier? Lastly, describe how you could achieve a non-linear decision boundary using a support vector classifier; give an example. What are potential drawbacks of this approach?

5. Consider a p-dimensional feature space $\mathcal{X}$, output space $\mathcal{Y} = \{-1, 1\}$ and a training set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. Define the concept of a support vector machine and explain all quantities involved. Give at least two non-trivial examples of various instances of this classifier. Describe the difference between a support vector machine and achieving non-linear decision boundaries using a support vector classifier. How does the cost parameter, C, in the definition of the support vector machine affect the resulting decision boundary? Lastly, describe one way a support vector machine can be used in the setting of $K > 2$ classes.

6. Suppose $\mathbf{x}$ comes from one of two populations, Π_1 with $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and Π_2 with $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. If the respective density functions are denoted by $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$. Find the expression for the quadratic discriminator Q, where
$$Q = \ln \left[\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \right].$$
If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, verify that Q becomes
$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2). \tag{2.5}$$

7. Assume that Π_i is $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$ and for the class prior probabilities π_i it hold that $\pi_1 = \pi_2 = \frac{1}{2}$. Specify the optimal misclassification rate (OMR) for the linear discriminant function. Explain further the relationship between the OMR and the Mahalanobis distance between two populations.

8. Relationship between LDA and QDA: Suppose that observations within each class are drawn from $N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $i = 1, \ldots, \mathcal{C}$, i.e from $\mathcal{C}$ classes with a class specific mean vector and a class specific covariance matrix. In general, as the sample size n increases, does one expect the test classification accuracy of QDA relative to LDA to improve, decline or be unchanged. Motivate your answer.

9. Suppose $\mathbf{x}_0$ comes from one of two populations, Π_1 with $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and Π_2 with $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Suppose also that you have a training data from both populations. Form the sample based linear discriminant function and specify the classification procedure for assigning $\mathbf{x}_0$ to one of two populations.

   What happens to Fisher's linear discriminant if the sample size n is smaller than the dimensionality p?

10. Explain the idea of k nearest neighbors, kNN, classifier. Derive the classification rule and explain the effect of k choice on the classification accuracy using both test and training errors.

11. State the multiple linear regression model and form the solution to the ordinary least-squares (OLS) parameter estimation problem. Assume further that the deviations of the response variable around its expectation are additive and Gaussian. Specify the distributional properties of the OLS estimators for this case and use them to form the test procedure for evaluation the partial effect of adding a covariate to the model. Add more assumptions if needed.

12. Explain the problem of collinearity in the linear regression model and suggest solutions.

13. The bootstrap approach provides an all-purpose method for computing standard errors. For any estimator $\hat{\theta} = \hat{\theta}(\mathbf{x})$ we can write

$$\mathrm{Var}^*(\hat{\theta}) = \frac{1}{n^n - 1} \sum_{i=1}^{n^n} \left(\hat{\theta}_i^* - \bar{\hat{\theta}}^* \right)^2,$$

   where * denotes a bootstrapped or re-sampled value, $\hat{\theta}_i^*$ is the estimator calculated from the i re-sample and $\sum_{i=1}^{n^n} \hat{\theta}_i^*$ is the mean of the re-sampled values. The variance formula above is applicable to virtually any estimator. But how do we know that it is a good method for evaluation the performance of estimators? Explain why the method is good using the following steps: for a sample $\mathbf{x} = (x_1, \ldots, x_n)$ and an estimate $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ select b bootstrap samples and calculate

$$\mathrm{Var}_b^*(\hat{\theta}) = \frac{1}{b - 1} \sum_{i=1}^{b} \left(\hat{\theta}_i^* - \bar{\hat{\theta}}^* \right)^2.$$

Explain why we have the convergence result

$$\mathrm{Var}_b^*(\hat{\theta}) \to \mathrm{Var}^*(\hat{\theta}) \text{ as } b \to \infty.$$

14. Consider the natural exponential family

$$f\left(x \mid \theta\right) = h(x)e^{\theta \cdot x - \psi(\theta)}.$$

Here $\theta \cdot x$ is inner product on R^k.

a) Consider the prior densities given by

$$\pi\left(\theta\right) = \psi\left(\theta|\mu, \lambda\right) = K\left(\mu, \lambda\right) e^{\theta \cdot \mu - \lambda\psi(\theta)}.$$

Find the the posterior as

$$\psi\left(\theta|\mu + x, \lambda + 1\right).$$

Hence the given family of priors is a conjugate family for $f\left(x \mid \theta\right)$. You are permitted to work formally, i.e., not checking the conditions for validity.

b) Let

$$f(x \mid \theta) = \prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i}; x_j \in \{0,1\}; 0 \leq \theta_i \leq 1.$$

$\theta = (\theta_1, \theta_2, \ldots \theta_d)$. Write this in a natural form and find the conjugate family of priors.

15. Suppose

$$P(B|A) > P(B).$$

We interpret this as the statement that if A is true then B is more likely. Formulate and prove the statements

If not-B is true, then A becomes less likely.

If B is true, A becomes more likely.

If not-A is true, then B becomes less likely.

16. We say that the discrete r.v.s X and Y are *conditionally independent* given Z if and only if for all x,y, z

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z).$$

Show that it follows from this definition that if X and Y are conditionally independent given Z, then

$$P(Y = y \mid X = x, Z = z) = P(Y = y \mid Z = z) \quad \text{for all } x,y,z.$$

17. In the text D. Poole & A. Mackworth: *Artificial Intelligence: Foundations of Computational Agents*, Cambridge University Press, 2010, we quote the following:

> *The idea of Bayesian learning is to compute the posterior probability distribution of the target features of a new example conditioned on its input features and all of the training examples.*
>
> *Suppose a new case has inputs $X = x$ and has target features, Y; the aim is to compute $P(Y|X = x \wedge e)$, where e is the set of training examples. This is the probability distribution of the target variables given the particular inputs and the examples. The role of a model is to be the assumed generator of the examples. If we let M be a set of disjoint and covering models, then reasoning by cases and the chain rule give*
>
> $$P(Y|x \wedge e) = \sum_{m \in M} P(Y \wedge m|x \wedge e)$$
>
> $$= \sum_{m \in M} P(Y|m \wedge x \wedge e) \times P(m|x \wedge e)$$
>
> $$= \sum_{m \in M} P(Y|m \wedge x) \times P(m|e).$$
>
> *The first two equalities are theorems from the definition of probability. The last equality makes two assumptions: the model includes all of the information about the examples that is necessary for a particular prediction (i.e., $P(Y|m \wedge x \wedge e) = P(Y|m \wedge x)$, and the model does not change depending on*

the inputs of the new example (i.e., $P(m|x \wedge e) = P(m|e)$. This formula says that we average over the prediction of all of the models, where each model is weighted by its posterior probability given the examples.

Obviously, the terminology above does not fully conform to the terminology made familiar in standard probability and statistics courses (abridged as sanstat below).

a) What does $\wedge$ represent in sanstat?

b) Verify the first two equalities by the rules of sanstat.

c) What is "reasoning by cases" in sanstat?

d) How would you interpret in sanstat "the model includes all of the information about the examples that is necessary for a particular prediction" ?

e) How would you interpret in sannstat "the model does not change depending on the inputs of the new example"?

d) What is the meaning of "the posterior probability distribution of the target features of a new example conditioned on its input features and all of the training examples"?

18. Let

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{elsewhere,} \end{cases}$$

where we know that $0 \leq \theta \leq 10$. We take for θ the prior density

$$\pi(\theta) = \begin{cases} \frac{1}{10} & 0 \leq \theta \leq 10 \\ 0 & \text{elsewhere.} \end{cases}$$

Asssume now that you have a data set $\mathcal{D}$ with four observations of x, and that $8 = \max \mathcal{D}$, i.e., 8 is largest value you have observed.

a) Show that the posterior distribution $\pi(\theta|\mathcal{D})$ is

$$\pi(\theta|\mathcal{D}) = \begin{cases} c \cdot \frac{1}{\theta^4} & 8 \leq \theta < 10 \\ 0 & \text{elsewhere,} \end{cases}$$

where you need not find the value of c.

b) Show that the predictive distribution

$$f(x \mid \mathcal{D}) = \int_0^{10} f(x|\theta)\pi(\theta|\mathcal{D})d\theta$$

is given by (c as in **a)**)

$$f(x \mid \mathcal{D}) = \begin{cases} \frac{c}{4} \cdot \left(\frac{1}{x^4} - \frac{1}{10^4}\right) & 8 \leq x < 10 \\ \frac{c}{4} \cdot \left(\frac{1}{8^4} - \frac{1}{10^4}\right) & 0 \leq x \leq 8 \\ 0 & \text{elsewhere.} \end{cases}$$

c) How would you comment the result in **b)**? Recall from first courses that the maximum liklelihood estimate of θ based on $\mathcal{D}$ is

$$\widehat{\theta}_{MLE} = 8$$

so the maximum likelihood method yields the probability density for future data

$$f(x \mid \widehat{\theta}_{MLE}) = \begin{cases} \frac{1}{8} & 0 \leq x \leq 8 \\ 0 & \text{elsewhere.} \end{cases}$$

19. Bayes Factor & Bayesian Model Comparison & Occam's Razor

Let

$$\mathcal{M}_i = \{p_i\left(x|\theta_i\right), \pi_i(\theta_i), \theta_0 \in \mathbf{\Theta}_i\}$$

be two Bayesian model families.

a) What is the marginal likelihood $p_i\left(x|\mathcal{M}_i\right)$ (of data) ?

b) How is the Bayes factor defined? Show that it can be written as

$$B_{01} = \frac{p_0\left(x|\mathcal{M}_0\right)}{p_1\left(x|\mathcal{M}_1\right)}.$$

c) What is *Occam's Razor*?

d) Suppose $\mathcal{M}_0$ is a more complex model family than $\mathcal{M}_1$, in the sense that $\mathcal{M}_0$ has more parameters. How does Bayes factor implement Occam's Razor?

20. Let $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ be a binary vector, i.e., $x_i \in [0,1], i \in \{1, \ldots, d\}$. There are two classes c_0 and c_1 and corresponding probability mass functions

$$p(\mathbf{x}|\theta_j, c_j) = \prod_{i=1}^{d} \theta_{ij}^{x_i}(1-\theta_{ij})^{1-x_i}; x_i \in \{0,1\}; 0 \leq \theta_{ij} \leq 1,$$

where $\theta_{ij} = P(\mathbf{X}_i = 1|c_j)$ and

$$\theta_j = (\theta_{1j}, \ldots, \theta_{dj}).$$

Then

$$p(c_j \mid \mathbf{x}) = \frac{p(\mathbf{x}|c_j)\pi_j}{C},$$

where $\pi_1 (= 1 - \pi_0)$ is the prior probability of c_1, and π_0 is the prior probability of c_0. C is the normalization constant that does not depend on c_1 and c_0.

Let Y be a r.v. such that

$$Y = \begin{cases} c_1 & \text{with probability } p(c_1 \mid \mathbf{x}) \\ c_0 & \text{with probability } p(c_0 \mid \mathbf{x}). \end{cases}$$

Show that Y follows a logistic regression and find $P(Y = c_1)$ in terms of the sigmoid function.

21. Let

$$Y^* = \boldsymbol{\beta}^T \mathbf{X} + \epsilon$$

where $\boldsymbol{\beta}^T \mathbf{X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$ and ϵ is a r.v. with the cumulative distribution function

$$P(\epsilon \leq x) = \frac{1}{1+e^{-x}} = \sigma(x).$$

Define Y as

$$Y = \begin{cases} 1 & \text{if } Y^* > 0 \text{ i.e. } -\varepsilon < \boldsymbol{\beta}^T \cdot \mathbf{X}, \\ -1 & \text{otherwise.} \end{cases}$$

Verify that Y follows a logistic regression w.r.t. $\mathbf{X}$.

22. Explain why the OLS parameter estimation in multiple regression model is not applicable when the dimensionality of the vector of independent variables, p exceeds the sample size, n. Suggest alternative estimation approaches and motivate why they work well for $p > n$ settings.

23. Explain the idea of the ridge regression and Lasso regression and the difference between these two approaches. Which of this two approaches behaves as a shrinkage method and which one can directly perform variable selection? Motivate your explanations by sketching the graph with profiles of ridge- and Lasso coefficient estimators as tuning parameter is varied, and explain the difference in profile shapes.

24. Suppose that $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$, where $\varepsilon_1, \ldots \varepsilon_n$ are independent and identically distributed with mean zero and constant variance σ^2. State the ridge- and Lasso problem in two forms, using the shrinkage penalty with parameter $\lambda \geq 0$ and using the explicit size constraint s on the parameter estimation. Explain the relationship between the parameters λ and s. Define the ridge and Lasso estimates of parameters $\beta_1, \ldots, \beta_p$.

25. Suppose that $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$, where $\varepsilon_1, \ldots \varepsilon_n$ are independent and identically distributed with mean zero and constant variance σ^2. The ridge coefficients minimize a penalized residual sum of squares

$$\hat{\beta}^{\text{ridge}} = \arg\min \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{i=j}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

where $\lambda \geq 0$ controls the amount of shrinkage. Rewrite this criterion in matrix form and show that the ridge regression solution can expressed as

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y},$$

where the input matrix $\mathbf{X}$ is centered, i.e. has p columns and $\mathbf{I}$ is the $p \times p$ identity matrix.

26. Explain the idea of k-fold cross-validation procedure in relation to estimation of the test and training error in the supervised learning. What are advantages and disadvantages of k-fold cross-validation relative to the validation set approach and leave-one-out cross-validation?

27. Explain the conceptual idea of parametric and non-parametric bootstrap and its application to quantifying uncertainty in statistical learning. Give examples. Suppose you have a sample $\mathbf{x} = (x_1, \ldots, x_n)$ and an estimator $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$. Derive the bootstrap based estimator of standard error, $\mathrm{SE}(\hat{\theta})$.

28. Let $\mathcal{X}^t = \left\{\mathbf{x}^{(l)}\right\}_{l=1}^{t}$, where $\mathbf{x}^{(l)} \in R^d$. be a training set. Let $D(\mathbf{x}, \mathbf{y})$ be a metric on $R^d \times R^d$.

a) Give in detail the steps of the k-means algorithm. What is the indata to the algorithm?

b) What is the definition of a Voronoi region, and how are Voronoi regions related to k-means clustering?

29. We have a statistical model that generates a set $\mathbf{X} = x$ of observed data, a set of unobserved latent data values $\mathbf{Z}$, and a vector of unknown parameters $\boldsymbol{\theta}$, along with a likelihood function

$$L(\boldsymbol{\theta}; \mathbf{X} = x, \mathbf{Z} = i) = P(\mathbf{Z} = i) p_{\mathbf{X}}(x|\boldsymbol{\theta}_i, \mathbf{Z} = i).$$

Hence $\boldsymbol{\theta}_i$ indicates the group of parameters in the conditional distribution of $\mathbf{X}$, when conditioning on $\mathbf{Z} = i$, and

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \boldsymbol{\theta}_k).$$

Let $L(\boldsymbol{\theta}; x)$ be the likelihood function when $\mathbf{Z}$ has been marginalized out.

$$L(\boldsymbol{\theta}; \mathbf{X} = x) = p_{\mathbf{X}}(x|\boldsymbol{\theta}) = \sum_{i=1}^{k} P(\mathbf{Z} = i) p_{\mathbf{X}}(x|\boldsymbol{\theta}_i)$$

The *EM- algorithm* (Expectation-Maximization -algorithm) seeks to find the MLE of the marginal likelihood $L(\boldsymbol{\theta}; x)$ by iteratively applying the following two steps:

Expectation step (*E step*): Calculate the expected value of the log likelihood function, with respect to the conditional distribution of $\mathbf{Z}$ given $\mathbf{X}$ under the current estimate of the parameters $\boldsymbol{\theta}^{(t)}$:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(t)}} \left[\log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})\right]$$

Maximization step (*M step*): Find the parameter that maximizes this quantity:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) .$$

Set $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t+1)}$, go to *E step* and continue until convergence.

This was the generic description. Consider now the situation where $k = 2$, $-\infty < \boldsymbol{\theta}_i < \infty$ for $i = 1, 2$, and

$$p_{\mathbf{X}}(x|\boldsymbol{\theta}_i, \mathbf{Z} = i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\boldsymbol{\theta}_i)^2}, \quad i = 1, 2, -\infty < x < \infty,$$

$P(\mathbf{Z} = 1) = \lambda_1$ and $P(\mathbf{Z} = 2) = \lambda_2 = 1 - \lambda_1$.

a) Assume that you have a training set of $\mathcal{X}^t = \{x^{(i)}\}_{i=1}^t$ regarded as t i.i.d. samples drawn from $f_{\mathbf{X}}(x|\boldsymbol{\theta})$. The corresponding values $\mathbf{Z}^{(i)}$ have not been observed. Find the Sundberg equations of the EM-algorithm for $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

b) Explain how this can be extended to a method of classification or unsupervised clustering and relate the method to k-means clustering. Why should this be called 'natural' k-means clustering?

30. The LDA is often taken as being synonymous with the term Fisher's linear discriminant. But Fisher's original study [1] of statistical discrimination actually establishes a different discriminant, which does not make some of the assumptions of LDA such as normally distributed classes or equal class covariances. We shall now take a look at the original concept. First some background.

[1] Fisher, R. A. (1936): The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics* 7 (2): 179−188.

Iris setosa

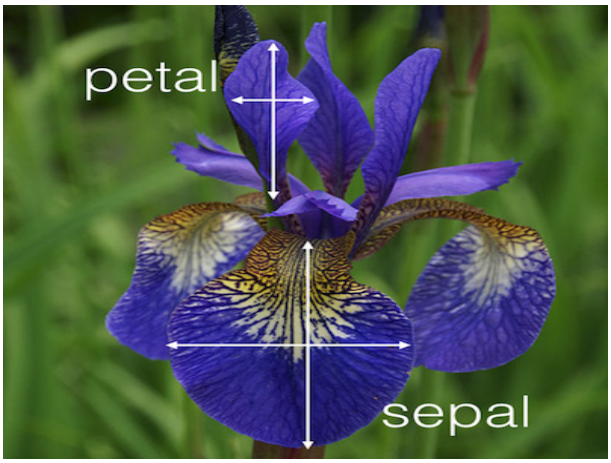
Iris versicolor

Iris virginica

Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. Two of the three species were collected in " all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus".

The data set consists of 150 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres. The data set can be downloaded at

`https://archive.ics.uci.edu/ml/datasets/Iris`



The Fisher discriminant invented for clustering/dealing with this data

set is based on *projection* of d-dimensional data *onto a line*. The hope is that these projections are separated by class. Thus the line is to be oriented to maximize this class separation. This line is taken to go through origin in R^d.

The $\mathbf{w}^T\mathbf{x}(=\sum_{i=1}^{d} w_i x_i)$ is the component of $\mathbf{x}$ in the direction of $\mathbf{w}$.

We try to find $\mathbf{w}$ so that the *separation criterion* (two classes)

$$J(\mathbf{w}) = \frac{\mu_1 - \mu_2}{\sigma_1^2 + \sigma_2^2}$$

is maximized, where $\underline{\mu}_i, i = 1, 2$ are two class centers in R^d, and

$$\mu_i = \mathbf{w}^T \underline{\mu}_i, \sigma_i^2 = \mathbf{w}^T \Sigma_i \mathbf{w},$$

and Σ_i is the covariance matrix of class i.

> Here the notation covers the cases, where $\underline{\mu}_i$ is known or is a class-wise sample mean
>
> $$\underline{\mu}_i = \begin{bmatrix} \bar{x}_{1i} \\ \vdots \\ \bar{x}_{di} \end{bmatrix} = \frac{1}{n} \sum_{j=1}^{n} x_{ji}$$
>
> and Σ_i is known or is a class-wise sample covariance matrix
>
> $$\Sigma_i = \frac{1}{n-1} \sum_{j=1}^{n} (x_{ji} - \underline{\mu}_i)(x_{ji} - \underline{\mu}_i)^{\mathrm{T}}.$$

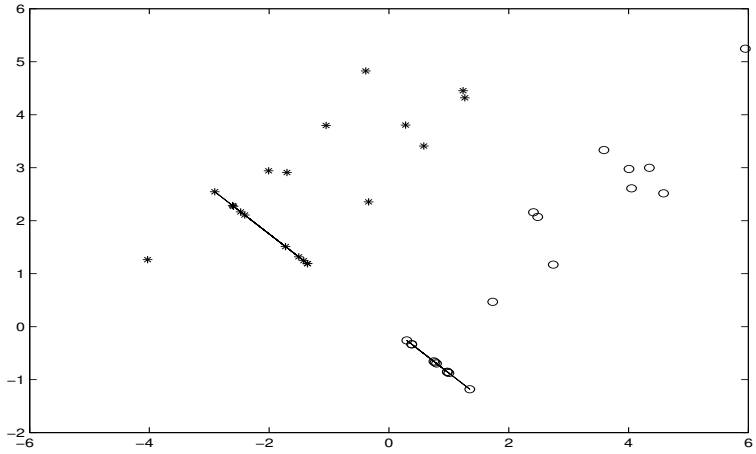
The criterion $J(\mathbf{w})$ can be maximized by standard methods and these give

$$\mathbf{w}^* = (\Sigma_1 + \Sigma_2)^{-1} \left(\underline{\mu}_1 - \underline{\mu}_2 \right), \tag{2.6}$$

and the optimal projection $P_{\mathrm{F}}(\mathbf{x})$ is

$$P_{\mathrm{F}}(\mathbf{x}) = \left(\mathbf{w}^{*^T} \mathbf{x} \right) \frac{\mathbf{w}^*}{\|\mathbf{w}^*\|^2} \tag{2.7}$$

A set of simulated two dimensional Gaussian data and their projections onto a line (oriented in Northwest-southeast direction, only segmentally depicted) are given in the figure below:



Fisher's linear discriminant can actually be arbitrarily bad ! However, it has far-reaching (kernelized) generalizations as support vector machines.

a) Is this an early (1936) example of a method of statistical learning? If it is, what sort of learning is being done? If it is not, where does the difference lie?
Hint: Be careful, the illustration of hierarchic clustering by Iris data in Lecture 17 need not lead here to the appropriate thinking.

b) Derive (2.6) and (2.7).
Aid: It is useful to try to do this by oneself, of course. If, however, external assistance is to be summoned up, a Google search with a phrase like *tutorial data reduction linear discriminant analysis* will very likely produce helpful documents.

c) Interpret the nature of the optimized $J(\mathbf{w}^*)$. Compare with (2.5) above.

31. We have four OTU's x, y, z, w in this order. We have computed the dissimilarity matrix

$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}$$

This means that, e.g., the dissimilarity between x and y is 0.3 and the dissimilarity between y and w is 0.8. Let $d_{i,j}$ denote an array in

this matrix. The hierarchical clustering method called single linkage (SLINK) works as follows:

i Find
$$(i^*, j^*) = \arg\min_{i,j} d_{i,j}$$

ii Join i^*, j^*

iii $d_{i,i^* \cup j^*} = \min(d_{i,i^*}, d_{i,j^*})$.

a) Use SLINK to cluster hierarchically the four OTU's and sketch the resulting dendrogram. Indicate on the plot the order in which the fusion occurs.

b) If we cut the dendrogram so that two clusters results, which OTU's are in each cluster ?

c) Check the ultrametric property on this dendrogram.

32. $\mathcal{P}$ is a finite set. The cardinality $\mid \mathcal{P} \mid$ of $\mathcal{P}$ is ≥ 2. The elements of $\mathcal{P}$ will denoted by $a, b, c, \ldots, p, \ldots, x, y$. The elements are abstract at this juncture, but are for the sake of definiteness to be called operative taxonomic units (OTUs).

- N is a finite positive integer.
- $M(n)$ is a function defined on integers n in $0 \ldots N$ and taking its values in partitions of $\mathcal{P}$, or, in the equivalence relations on $\mathcal{P}$ such that the 'Woodger-Gregg' axioms(a)-(c) hold:
  (a) $M(0) = \{(p, p) \mid p \in \mathcal{P}\}$.
  (b) $M(N) = \mathcal{P} \times \mathcal{P}$.
  (c) $0 \leq n \leq m \leq N \Rightarrow M(n) \subseteq M(m)$.

We shall denote the set of equivalence classes in the image of $M(n)$ by

$$\mathbf{P}(n) = \mathcal{P}/M(n). \tag{2.8}$$

Each $\mathbf{P}(n)$ is a partition of $\mathcal{P}$.

From this and from the properties (a)-(c) it follows that for $n \leq m$

$$\mathbf{P}(0) \leq \mathbf{P}(n) \leq \mathbf{P}(m) \leq \mathbf{P}(N),$$

where $\leq$ is the usual order for partitions, i.e., $\mathbf{P}(n) \leq \mathbf{P}(m)$ means that $\mathbf{P}(m)$ is a coarser partition of $\mathcal{P}$ than $\mathbf{P}(n)$. $\mathbf{P}(N)$ is the maximal element and $\mathbf{P}(0)$ is the minimal element, and $\mathbf{P}(N)$ is the coarsest, where all OTU's are in a single set. $\mathbf{P}(0)$ is the the most refined partition consisting of singletons. Clearly $\leq$ is a partial order and $(\mathbf{P}(n))_{n=0}^{N}$ is a lattice. Now we introduce a technical conceptual aid, which is, called taxon in a hierarchy.

Definition 2.1 *Y is a taxon* in the hierarchy $(\mathcal{P}, N, M)$, if there is an integer n in $0, \ldots, N$ such that

$$Y \in \{(A, n) \mid A \in \mathbf{P}(n)\}.$$

■

Here (A, n) is an ordered pair. If Y is a taxon in a hierarchy, we say that

$$\mathrm{Ext}(Y) = A, \mathrm{Rank}(Y) = n.$$

On occasion we shall also talk about the sets $\mathrm{Ext}(Y)$ as extensions. In this terminology, $\mathbf{P}(n) \leq \mathbf{P}(m)$ means that every extension of a taxon in $\mathbf{P}(m)$ is a union of extensions in $\mathbf{P}(n)$.

Next we can introduce a map $J(x, y)$ from pairs of OTU's to $\{0 \ldots N\}$.

Definition 2.2 Let $x, y \in \mathcal{P}$ and let Y be the taxon of lowest rank, $\mathrm{Rank}(Y)$, in the hierarchy $(\mathcal{P}, N, M)$ such that $x, y \in \mathrm{Ext}(Y)$. Then we set

$$J(x, y) = \mathrm{Rank}(Y). \tag{2.9}$$

■

The definition presupposes that Y is unique, which holds by the properties of partitions. Due to this we can introduce $\mathrm{lca}(x, y)$, the *least common ancestor* of x and y by

$$\mathrm{lca}(x, y) = \mathrm{Ext}(Y), \quad \text{such that} \quad \mathrm{Rank}(Y) = J(x, y). \tag{2.10}$$

- Show that $J(x, y)$ is an ultrametric on $\mathcal{P} \times \mathcal{P}$.

33. Describe the idea of CLINK in hierarchical clustering.

34. Describe how to build a regression or decision tree using cost complexity pruning. When building a classification tree using a cost function based on, e.g., the Gini index or the cross-entropy, it may happen that a split leaves two terminal nodes with the *same* predicted class. How is this possible?

35. In the random forest approach, every node split considers cost function minimization with respect to only a subset of $m \leq p$ randomly chosen predictors (where p denotes the total number of predictors). Show that the probability that a given predictor is not considered at a given split is $(p-m)/p$. What is the rationale behind considering only a randomly chosen subset of the predictors?