Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

EXAM IN
MA8701 GENERAL STATISTICAL METHODS

Tuesday May 16, 2017
09:00 – 13:00

Aids: *Simple calculator (NTNU)*

Problem 1 *Supervised learning*

a) What is meant by *supervised learning*?

What are the main objectives of supervised learning?

Formalize briefly the problem of supervised learning in, respectively, a *regression setting* and a *classification setting*.

b) In supervised learning it is common to divide the data set in a *training set* and a *test set*, and sometimes also a *validation set*.

Discuss the role of these sets and point to advantages/disadvantages of making such a division of the data set.

c) Explain what is meant by *cross-validation*. Discuss its use in practice.

How does cross-validation relate to the use of training/validation/test sets?

d) *Bootstrap* methods are often used as an alternative to cross-validation. Explain briefly how a bootstrap-estimate of prediction error can be found.

e) Bootstrapping is also used in *bagging*. What is the main idea here? What is the connection between bagging and *random forests*?

f) What is meant by *curse of dimensionality*? Why is the curse of dimensionality particularly problematic for k-nearest-neighbor methods?

Problem 2 *Regression trees*

Consider training data $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$, where the $\mathbf{x}_i$ are in R^p while the y_i are real numbers. The problem to be considered is that of growing a *regression tree* based on these data.

a) Explain briefly what is meant by a regression tree, and explain in particular how it is constructed by using *recursive binary splittings*.

In this connection, discuss briefly the advantages by letting each single split depend on only one of the input variables.

b) The *cost-complexity criterion* for pruning of a regression tree can be written as

$$C_\alpha(T) = \sum_{m=1}^{|T|} \sum_{\mathbf{x}_i \in R_m} (y_i - \hat{c}_m)^2 + \alpha|T|. \tag{1}$$

Explain the ingredients of this expression, and how it is used in the process of *cost-complexity pruning*.

Suppose in the following that $p = 2$ and $N = 4$ and consider the toy training data given by

i	(x_{i1}, x_{i2})	y_i
1	$(1, 3)$	2
2	$(2, 2)$	5
3	$(3, 2)$	3
4	$(3, 4)$	7

c) Find the optimal splitting variable and split point for the first binary splitting for these data. Indicate how you use the general algorithm to do this.

(*Hint:* Draw a figure and look at possible divisions).

d) Continue the tree construction for the toy data until each node in the tree corresponds to a single observation. Let the resulting tree be denoted T_0. For what values of α in the cost-complexity criterion (1) will the unprunded tree T_0 be the optimal tree?

e) Suppose that in the toy example one wants to predict the response y for a new observation at $\mathbf{x} = (2, 3)$.

What is the predicted value when using the tree T_0 constructed above?

What would be the prediction if one used the 2-Nearest-Neighbor method? (Explain how you get the result here).

What is the predicted value if an ordinary linear regression model is fit to the data? You need not do the complete calculation here, just indicate how you would proceed.

Problem 3 *Classification*

Let $(\mathbf{X}, G)$ be a random pair where the input $\mathbf{X}$ is a random vector in $\mathbb{R}^p$ and G is a categorical variable, $G \in \{1, 2, \ldots, K\}$, denoting the class from which the observation $\mathbf{X}$ comes. The task is to predict the class G from an observation of $\mathbf{X}$, with loss given by 0 for a correct classification and 1 for a wrong classification.

a) Show that the optimal strategy for an observed $\mathbf{X} = \mathbf{x}$ is to classify to the class g which maximizes $P(G = g|\mathbf{X} = \mathbf{x})$.

This strategy is called the *Bayes classifier*, and the corresponding expected prediction error is called the *Bayes rate*. Find an expression for the latter.

Let the joint distribution of $(\mathbf{X}, G)$ be given by $f(\mathbf{x}, g)$, which is assumed to be a density in $\mathbf{x} \in \mathbb{R}^p$ and a point mass in $g \in \{1, 2, \ldots, K\}$. There are two ways of representing this joint distribution of $(\mathbf{X}, G)$:

(i) $f(\mathbf{x}, g) = f(\mathbf{x}|G = g)\pi_g$

(ii) $f(\mathbf{x}, g) = P(G = g|\mathbf{X} = \mathbf{x})f_{\mathbf{X}}(\mathbf{x})$

b) *Linear Discriminant Analysis* (LDA) is based on one of these representations. Which one, and which assumptions are made for the ingredients of the right hand side of the representation?

Logistic regression is also based on one of these representations. Which one, and which assumptions are made for the ingredients of the right hand side of the representation?

c) Suppose that the distribution of (X, G) is unknown but that there are given training data $(x_1, g_1), (x_2, g_2), \ldots, (x_N, g_N)$ drawn from the distribution of (X, G).

How would you use these data to derive a classifier in each of the two cases considered in subpoint (b) above? (Explain very briefly).

d) Let $p = 1$ and $K = 2$ and let the distribution of X for given class $G = g$ be exponential with hazard rate λ_g, i.e., suppose that

$$f(x|G = g) = \lambda_g \exp(-\lambda_g x)$$

for $x > 0$, $g = 1, 2$, where $\lambda_1, \lambda_2 > 0$. Let further $\pi_g = P(G = g)$ for $g = 1, 2$.

Find the Bayes classifier.

Consider in particular the special case when $\lambda_1 = 2$, $\lambda_2 = 1$ and the two classes are equally probable. What is the Bayes rate in this case?