Name:.....

Data Mining and Visualization

Assignment 3

Classification and logistic regression models

Assignments constitute part of the examination and must be handed in time. You are asked to hand in the solutions during a week following the week on which the assignment has been discussed in classes. You can submit either an electronic copy or a hard copy of your work. In the latter case, staple your solutions together.

Problem 1 – Additive logistic regression

The problem of classification was discussed in the lecture and logistic regression was proposed as a model for it. Let us consider classification into two groups. Check the lecture to provide answers to the following questions that are based on an analysis of some real data. The data are a subset of the Coronary Risk-Factor Study (CORIS) baseline survey, carried out in three rural areas of the Western Cape, South Africa. The aim of the study was to establish the intensity of ischemic heart disease risk factors in that high-incidence region. The data represent white males between 15 and 64, and the response variable is the presence or absence of myocardial infarction (MI) at the time of the survey (the overall prevalence of MI was 5.1% in this region). A logistic-regression model has been fit by the maximum likelihood method and summarized in the following table (only significant factors have been left). This summary includes Z scores for each of the coefficients in the model (coefficient estimates divided by their standard errors). The factors are: family history, cholesterol levels (ldl), age, tobacco use.

	Coefficient	Std. Error	Z score
`(Intercept)`	−4.204	0.498	−8.45
`tobacco`	0.081	0.026	3.16
`ldl`	0.168	0.054	3.09
`famhist`	0.924	0.223	4.14
`age`	0.044	0.010	4.52

Knowing that the linear logistic regression has been used to model the data answer the following.

1. Write down explicitly formula for the probability that a person will have the myocardial infarction (heart attack).
2. Suppose that a person have the following values for the four factors: family history - YES, cholesterol level - 6, age - 58, tobacco use - 10. What are his chances to have the heart attack?
3. If one consider the high risk group as those who has the chances of the heart attack higher than 50%. Provide with a classification rule for ruling a person to be in the high risk group.
4. Would an individual at age 45, with tobacco use - 20, family history - NO, cholesterol level - 8, be in the high risk group?

Problem 2 – Bayes theorem

Consider a logistic regression model with Y taking values zero or one and with two continuous predictors x_1 and x_2. Assume that the predictors are distributed in the population as jointly normal random variables correlated with each other with means 40 and 3.5, respectively and standard deviations of 10% of the means, while correlation coefficient $\rho = 0.7$.

Let us now consider a classification rule: $\mathcal{X}_1 = \{(x_1, x_2) : P(x_1, x_2) > 0.95\}$.

Answer the following questions

1. Give simple rationale behind using such a rule.
2. Formulate the Bayes theorem and using it derive the formula for the misclassification rates as given in the lecture.
3. Under the above assumptions about the model write explicitly the classification rule in terms of the parameters α, β_1 and β_2 in the logistic regression.
4. Show that the misclassification rates depend on $P(x_1, x_2)$ through

$$\frac{P(Y = 0|\alpha + \beta_1 x_1 + \beta_2 x_2 < 2.944439)}{P(Y = 0|\alpha + \beta_1 x_1 + \beta_2 x_2 \geq 2.944439)}$$

and

$$\frac{P(Y = 1|\alpha + \beta_1 x_1 + \beta_2 x_2 \geq 2.944439)}{P(Y = 1|\alpha + \beta_1 x_1 + \beta_2 x_2 < 2.944439)}.$$

5. Consider $\alpha = 1$, $\beta_1 = 0.02$ and $\beta_2 = 0.3$ and propose a method of evaluation the misclassification errors (it can be based on the Monte Carlo method). How, will they change if $\alpha = 0.75$, $\beta_1 = 0.1$ and $\beta_2 = -0.8$?

Problem 3 – Properties of additive logistic likelihood

In the lecture on the generalized additive models the formula for the logistic additive model log-likelihood and its first and second derivative have been presented. Here, we ask for the mathematical details of their derivation.

1. Recall that the log-likelihood has the form

$$\ell(\alpha, \beta) = \sum_{i=1}^{N} y_i(\alpha + f_1(X_{i1}) + \cdots + f_p(X_{ip})) - \log(1 + e^{\alpha + f_1(X_{i1}) + \cdots + f_p(X_{ip})}).$$

   Show that if $f_j(x) = \beta_j x$, then its partial first order derivatives are

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{N} (y_i - p(\mathbf{x}_i, \alpha, \beta)),$$
$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{N} x_{ij}(y_i - p(\mathbf{x}_i, \alpha, \beta)), \quad j = 1, \ldots, p$$

2. Then the matrix of the second partial derivatives is

$$\frac{\partial^2 \ell(\alpha, \beta)}{\partial(\alpha, \beta)\partial(\alpha, \beta)^T} = -\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T p(\mathbf{x}_i; \alpha, \beta)(1 - p(\mathbf{x}_i; \alpha, \beta)),$$

   where $\mathbf{x}_i = (1, x_1, \ldots, x_p)$ is treated as a column $p \times 1$ matrix.

Problem 4 – Mixture, likelihood ratio and two classes with prior probabilities

The problem of classification was discussed in the lecture and logistic regression was proposed as a model for it. Let us consider classification into two groups. Suppose that prior to obtaining information about some features, one knows that in general population there are twice as many individuals in Class 0 as in Class 1. Now, we measure some variables, say X_1,X_2,X_3 and they have distribution given by density $f(x_1, x_2, x_3|0)$ among individuals in Class 0 and by density $f(x_1, x_2, x_3|1)$ among individuals in Class 1.

The posterior chances of being in Class 0 after observing values x_1, x_2, x_3 for three variables are

$$P(0|x_1, x_2, x_3) = \frac{f(x_1, x_2, x_3|0) * 2/3}{f(x_1, x_2, x_3|0) * 2/3 + f(x_1, x_2, x_3|1) * 1/3}$$

and for being in Class 1 are

$$P(1|x_1, x_2, x_3) = \frac{f(x_1, x_2, x_3|1) * 1/3}{f(x_1, x_2, x_3|0) * 2/3 + f(x_1, x_2, x_3|1) * 1/3}.$$

The model like above is often referred to as the two component mixture model.

1. Suppose that variables X_1, X_2, and X_3 are independent and exponentially distributed with parameters β_1, β_2 and β_3, and λ_1, λ_2, λ_3, for Class 0 and Class 1, respectively. Write explicitly the model for posterior probabilities. Is the model based on the posterior resembling linear logistic regression? What is similar and what is different?
2. Consider that the data X_1, X_2, and X_3 are observed independently from a gamma distribution (density is proportional to $x^{\tau-1}e^{-\beta x}$) with different shape parameters τ but with mean equal to one (so you have two sets of three τ's). Derive the classification rule for this case, assuming that the prior probabilities are equal. Is this model similar to the linear logistic model?
3. How the additive model can be related to the model based on the posterior distributions in the discussed case?

Problem 5 – Additive logistic regression vs two component mixture Review from the lecture the additive logistic regression model in the context of classification and the two component mixture model.

1. Consider that the data are coming from the two component mixture model that has one dimensional independent normally distributed components. Write the true relation between the feature variable X and the logit function.
2. What would be the function $f(X)$ in the additive logistic regression model to have the structure of this normal mixture? Under which assumptions about the parameters of the components the logistic model is linear?
3. Consider now that the variable $X = (X_1, X_2)$ is two dimensional and distributed as a mixture of the two dimensional normal random variables with some correlation. Can the two component mixture model in this case can be viewed as an additive logistic regression model? Can some special assumptions about the parameters lead to the additive or linear logistic regression model?