

# STATS 211: Statistical Methods II

## Lecture 9: Classification models

Babak Shahbaba

UCI, Winter 2012

## Classification models

- In this lecture, we discuss situations where the response variable,  $y$ , is categorical, e.g., diseased vs. healthy, mutant gene vs. wild-type, spam vs. no-spam, different forest cover types, different types of lung cancer, etc.
- In these situations, our main goal is either investigating the relationship between  $y$  and a set of explanatory variables  $x$ , or classifying future observations into one of the possible categories. This course is concerned with the latter.
- For these models, the normality assumption commonly used in standard linear regression models is not appropriate.
- We could of course use generalized linear models (GLM) such as logistic (for binary response variables) and multinomial logistic (for multiple categories) regression models.
- In this course, however, we focus on classification models using discriminant analysis and naive Bayes classifiers.

## Classification models

- Suppose each case can belong to one of  $K$  possible categories; that is,  $y$  can take a value from  $1, 2, \dots, K$ .
- In order to classify cases to one of the possible categories based on the observed values of their covariates  $x$ , we need to obtain a set of *discriminant functions*,  $\delta_1(x), \dots, \delta_K(x)$ , such that we set  $\hat{y} = k$  if its corresponding discriminant function  $\delta_k(x)$  has the highest value among all discriminant functions.
- For some classification models (e.g., logistic regression models and naive Bayes classifiers),  $\delta_k(x)$  are set to the class probabilities or any monotonically increases function of these probabilities.
- In some other classification models (e.g., support vector machines),  $\delta_k(x)$  are not directly related to class probabilities.

## Discriminative vs. generative

- Models that define discriminant functions based on class probabilities can be divided into *discriminative* and *generative* groups.
- Discriminative models estimate the conditional distribution  $P(y|x)$ , but not the distribution of covariates,  $P(x)$ .
- *generative* models estimates the joint distribution of response and covariates,  $P(x, y)$ .
- The joint distribution, of course, can be used to find the conditional distribution  $P(y|x)$  using Bayes' rule:

$$P(y = k|x) = \frac{P(y = k)P(x|y = k)}{P(x)}$$

- Throughout this course,  $P(x)$  denotes the probability mass function when  $x$  is discrete, and it denotes the density function for continuous variables.

## Discriminative vs. generative

- Generative models have several advantages over discriminative models. They provide a natural framework for handling missing data or partially labeled data. They can also augment small quantities of expensive labeled data with large quantities of cheap unlabeled data. This is especially useful in applications such as document labeling and image analysis, where it may provide better predictions for new feature patterns not present in the data at the time of training.
- While generative models are quite successful in many problems, they can be computationally intensive. Moreover, finding a good (but not perfect) estimate for the joint distribution of all variables (i.e.,  $x$  and  $y$ ) does not in general lead to good predictions.
- By contrast, discriminative models are often computationally fast and are preferred when the covariates are in fact non-random (e.g., they are fixed by an experimental design).

## Linear discriminant analysis

- When the set of  $p$  covariates,  $x$ , are continuous random variables, we can assume that their joint distribution is multivariate normal for each class,

$$f_k(x) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k)\right]$$

- Note that only the mean of the distributions,  $\mu_k$ , changes from one class to another. The covariance matrix  $\Sigma$  remains the same for all classes.
- This assumption is of course not realistic and is made only for simplicity. We will relax it later.

## Linear discriminant analysis

- Using Bayes theorem, we have

$$P(y = k|x) = \frac{\pi_k f_k(x)}{\sum_{k'=1}^K \pi_{k'} f_{k'}(x)}$$

where  $\pi_k = P(y = k)$ .

- For a given value of  $x$ , the denominator remains the same for all classes. Therefore, we can define the discriminant function based on the numerator,  $\pi_k f_k(x)$ , or more commonly based on its log,

$$\begin{aligned}\delta_k(x) &= \log \pi_k + \log[f_k(x)] \\ &= \log \pi_k - \frac{1}{2} \log(|\Sigma|) - \frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k)\end{aligned}$$

## Linear discriminant analysis

- With further simplification, we have

$$\delta_k(x) = \log \pi_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + x^T \Sigma^{-1} \mu_k$$

- Note that the above functions are linear in  $x$ .
- Therefore, we refer to them as the *linear discriminant functions*.
- Classifying cases according to these functions is called the *linear discriminant analysis* (LDA).

## Linear discriminant analysis

- We can estimate  $\pi_k$  and  $\mu_k$  for  $k = 1, \dots, K$ , and  $\Sigma$  as follows:

$$\hat{\pi}_k = \frac{n_k}{n}$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k}^{n_k} x_i$$

$$\hat{\Sigma} = \frac{1}{n - K} \sum_{k=1}^K \sum_{i:y_i=k}^{n_k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

where  $n_k$  is the number of observed cases (training cases) that belong to class  $k$ .

## Linear discriminant analysis

- After estimating the model parameters, we assign each case,  $i$ , to the class whose value of the discriminant function,  $\delta_k(x_i)$ , is the highest.
- Cases for which  $\delta_k(x) = \delta_l(x)$  fall on the decision boundary between the two classes  $k$  and  $l$ .
- For these cases,  $\delta_k(x) - \delta_l(x) = 0$ , which means

$$\log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k - \mu_l)^T \Sigma^{-1}(\mu_k - \mu_l) + x^T \Sigma^{-1}(\mu_k - \mu_l) = 0$$

- Note that the above equation, which specifies the decision boundary, is linear in  $x$ . As the result, the decision boundaries *hyperplanes* in  $p$  dimensions. (The decision boundary is straight line if we have two covariates only.)

## Quadratic discriminant analysis

- As mentioned above, the equal-covariance assumption is restrictive and is only made for convenience.
- By relaxing this assumption, the discriminant function becomes

$$\delta_k(x) = \log \pi_k - \frac{1}{2} \log(|\Sigma_k|) - \frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)$$

which are quadratic functions of  $x$ ; hence, called *quadratic discriminant functions*.

- Classifying cases according to these functions is called *quadratic discriminant analysis* (QDA).
- The decision boundaries for this approach are not linear any more.

## Naive Bayes models

- This is an alternative classification model, which is especially attractive when the dimension  $p$  is large.
- In this approach, we use Bayes theorem to obtain the probability of each class given the observed values of covariates,

$$P(y = k | x_1, \dots, x_p) = \frac{P(y = k)P(x_1, \dots, x_p | y = k)}{\sum_{k'=1}^K P(y = k')P(x_1, \dots, x_p | y = k')}$$

- Now, we make an assumption that is naive and possibly wrong, but it simplifies the model; we assume that given a class  $y = k$ , the covariates are independent,

$$P(x_1, \dots, x_p | y = k) = \prod_{j=1}^p P(x_j | y = k)$$

## Naive Bayes models

- As the result of the above naive assumption, the model simplifies as follows:

$$P(y = k | x_1, \dots, x_p) = \frac{P(y = k) \prod_{j=1}^p P(x_j | y = k)}{\sum_{k'=1}^K P(y = k') \prod_{j=1}^p P(x_j | y = k')}$$

- As before, we assign each case,  $i$ , to the class with the highest conditional probability given  $x_{i1}, \dots, x_{ip}$ .
- It is more common to distinguish between two classes using the following logit function

$$\begin{aligned} \log \frac{P(y = k | x_1, \dots, x_p)}{P(y = l | x_1, \dots, x_p)} &= \log \frac{P(y = k) \prod_{j=1}^p P(x_j | y = k)}{P(y = l) \prod_{j=1}^p P(x_j | y = l)} \\ &= \log \frac{\pi_k}{\pi_l} + \sum_{j=1}^p \log \frac{P(x_j | y = k)}{P(x_j | y = l)} \end{aligned}$$

## Naive Bayes models

- In practice, we estimate  $\pi_k$  using the proportion of observed cases that belong to class  $k$ .
- To estimate  $P(x_j|k)$ , we first need to assume a probability distribution model for  $x_j$  given  $k$ .
- If  $x_j$  is categorical, we can estimate  $P(x_j|k)$  using the observed proportion of each category of  $x_j$  for cases with  $y = k$ .
- If  $x_j$  is continuous, we can assume  $x_j|k$  has a Gaussian distribution and estimate its mean and variance using the cases with  $y = k$ .

## Predictive power

- After we build classifiers, we evaluate their performance by measuring their predictive power.
- A common measure for predictive power is *accuracy rate*, which is defined as the percentage of the times the correct class is predicted for future observations (or observations in the test set).

$$acc = \frac{\sum_{i=1}^{n_t} I(\hat{y}_i = y_i)}{n_t}$$

where  $n_t$  is the number of observations in the test set,  $y_i$  is the true class, and  $\hat{y}_i$  is the predicted class for  $i^{th}$  observation in the test set. The index  $i$  here is for test cases.

- Instead of accuracy rate, we could also use error rate, which is defined as the percentage of the times the wrong class is predicted.

## Predictive power

- Instead of averaging over all predictions, it might be more informative to separate the types of error.
- One common approach for doing this is to present the results in a *classification table*.
- For classification problems with two possible classes,  $\{0, 1\}$ , we have

		Predicted class	
		0	1
True class	0	True Negative	False Positive
	1	False Negative	True Positive

- Based on this table, we have

$$\text{Sensitivity} = P(\hat{y} = 1 | y = 1)$$

$$\text{Specificity} = P(\hat{y} = 0 | y = 0)$$

## Evaluating performance

- When there are multiple classes, we can use the  $F_1$  measure instead,

$$F_1 = \frac{1}{K} \sum_{k=1}^K \frac{2A_k}{2A_k + B_k + C_k}$$

- Here,  $A_k$  is the number of cases which are correctly assigned to class  $k$ .
- $B_k$  is the number cases incorrectly assigned to class  $k$ .
- $C_k$  is the number of cases which belong to the class  $k$  but are assigned to other classes.
- The higher the  $F_1$  measure the better the model.