Classification – Fundamentals and Overview

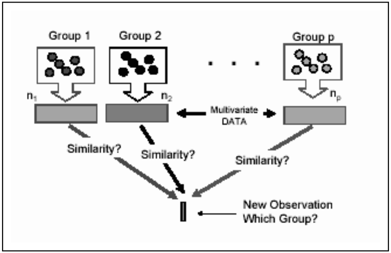
September 17, 2019

Classification goal

- **Overall goal**: We observe certain **features of an object** and we want decide to which **category** (or class, or population) this object belongs.
- The classification of an object to a class is made through a classification rule.
- **Goal**: Find an effective classification rule.

Discrimination, validation, and testing

- **Discriminate** between classes, i.e. identify relevant features for the classification problem and propose models and methods that allow to develop reasonable classification rules – **learning phase**
- **Verify** how these methods perform on actual data sets and decide for the optimal method
- **Test** how the optimal method performs on a data set that was not used for the discrimination and method selection stages.



Data allocation – data mining approach

- Allocate data, for example 50% for **the learning phase** (training), 25% for **validation** (model/method selection), and 25% for **the testing phase** (final model assessment)
- **Training**: using data to propose a number/class of possible models that maybe adequate.
- **Model/method selection**: estimating the performance of different models or methods in order to choose the best one.
- **Final model assessment**: having chosen a final model, estimating its prediction error on 'fresh' testing data.

Few examples

- A scientist needs to **discriminate between earthquake and an underground nuclear explosion** on the basis of signals recorded at a seismological station.
- An economist wishes to **forecast on the basis of accounting information** those members of the corporate sector that might be expected to suffer financial **losses leading to a bankruptcy**.
- A veterinarian has information on **the age, weight and radiographic measurements** for three groups of dogs: **Normal healthy, Bowel obstructed, Chronic diseased**.

  A dog enters the clinic and its age, weight and radiographic measurements are determined. To which group should it be classified?
- Automatic spam detector – **predicting (classifying) whether the email was junk email.**
- Using some available sociometric information extracted from social networks **predict that an individual's income exceeds $250, 000 per year.**

Notation

- An object with features' measurement $\mathbf{X}$: $p \times 1$ vector. It belongs to one of two classes **0** or **1**.
- A **selection rule** is a split of the feature space into two parts $\mathcal{X}_0$ and $\mathcal{X}_1$.
  - If $\mathbf{x} \in \mathcal{X}_0$ classify to class **0**.
  - If $\mathbf{x} \in \mathcal{X}_1$ classify to class **1**.
- $Y = 0$ if the object at hand is in class **0** and $Y = 1$ if in class **1**.
- Y is not observed, in general, but the values of Y are known for training, validation, and test data.
- Classification as a **prediction binary variable**:

$$R(\mathbf{X}) = \begin{cases} 1; & \mathbf{X} \in \mathcal{X}_1 \\ 0; & \mathbf{X} \in \mathcal{X}_0 \end{cases}$$

- R is dependent entirely on $\mathbf{X}$ so it is random only if $\mathbf{X}$ is random but in any case if $\mathbf{X}$ is known, then $R(\mathbf{X})$ is known too.

Formulation of the problem

- **Goal:** Make R as close as possible to Y (if R is equal to Y then the prediction/classification is perfect).
- $Y = 1$ or $Y = 0$ – Y a binary variable (outcome)
- $\mathbf{X} = (X_1, \dots, X_p)$ – predictor, features
- The chances that the object with features **X** is in the class **1** can be viewed as the conditional probability given **X**:

$$P(\mathbf{X}) = P(Y = 1|\mathbf{X}) = P(X_1, \dots, X_p)$$

- Features can be viewed random or not. If they are not random the above is considered as a probability dependent on features.
- If they are viewed random the classification rule can exploit their random distributions.

How to define R (to decide for regions $\mathcal{X}_0$ and $\mathcal{X}_1$)?

Three major approaches based on probability:

- Use **binomial likelihoods** for Y given that $\mathbf{X}$ are non-random, this was discussed before as the **logistic regression**:

$$\log \frac{P(Y=1|X_1,\ldots,X_p)}{P(Y=0|X_1,\ldots,X_p)} = \alpha + f_1(X_1) + \cdots + f_p(X_p)$$

- Use **likelihoods for X if one can consider them X to be random** – the binary value of Y gives a choice of parameters for the distribution of $\mathbf{X}$:

$$g(\mathbf{x}|Y=1) = g_1(\mathbf{x})$$
$$g(\mathbf{x}|Y=0) = g_0(\mathbf{x})$$

  The **likelihood ratio** with **estimated parameters** can be used to define a classification rule.

- Assume **prior distribution for** Y treat $\mathbf{X}$ as random and use **posterior probabilities for Y** to define a classification rule– **Bayesian approach**.

Logistic regression vs. posterior distributions

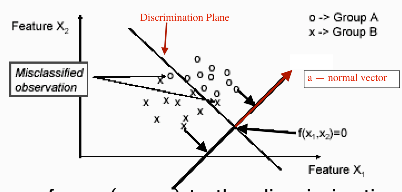
- The first two approaches are, in fact, connected, see Assignment 3. Namely, **additive logistic regression** can be viewed as a **likelihood approach** with assumed independence between features X_i's.
- The main conceptual difference in the approaches is that in the second approach **explanatory variables** X (features) are considered **random** and some concrete models for their probability distribution can be imposed.
- The **posterior distribution** approach assumes some parametric structure for distributions of variables X_i's plus some **prior chances** for membership in the classes.

The approaches are related through Bayes theorem relation

$$P(Y = 1|X_1, \dots, X_p) \sim P(X_1, \dots, X_p|Y = 1)P(Y = 1).$$

Geometric approach – without any probability

- For the training data find a **discrimination plane** that the best divides between two groups. Let $\mathbf{a}$ be any vector that is perpendicular to this plane.



Let $\mathbf{Px}$ be the projection of $x = (x_1, x_2)$ to the discrimination plane and $\mathbf{a}$ is any vector perpendicular to it, decide for Group A if

$$f(x_1, x_2) = (\mathbf{x} - \mathbf{Px})^T \mathbf{a} = \mathbf{x}^T \mathbf{a} > 0$$

and Group B otherwise. In the above we used that $\mathbf{Px}^T \mathbf{a} = 0$. Why is it true?

- Note that $f(x_1, x_2) = \|\mathbf{a}\| \|\mathbf{x}\| \cos \alpha$, where α is the angle between $\mathbf{a}$ and $\mathbf{x}$, so we decide for the membership based if the angle is greater or smaller than $\pi/2$.
- **How good is such a classification rule?**

Misclassification probabilities with prior distribution

- The observations are coming from the two classes according to the **prior distribution** given by $p_0 \in [0,1]$ and $p_1 = 1 - p_0$, i.e. $Y = 0$ if the object in hand is in Class **0** and $Y = 1$ otherwise (Class **1**) and

$$P(Y=0) = p_0, \; P(Y=1) = p_1 = 1 - p_0$$

- Given that the observation is from Class **0** the chance for it to be **misclassified** is denoted by $P(1|0) = P(R=1|Y=0)$ and analogously if it comes from Class **1** the chance for it to be misclassified is denoted by $P(0|1) = P(R=0|Y=1)$.

$$\begin{aligned} P(\text{Error}) = & P(R=0|Y=1)P(Y=1) + P(R=1|Y=0)P(Y=0) = \\ & = P(0|1)p_1 + P(1|0)p_0 \end{aligned}$$

- **Expected cost of misclassification**: $c(0|1)$, $c(1|0)$ stand for the respective costs of misclassification:

$$\text{ECM} = c(0|1)P(0|1)p_1 + c(1|0)P(1|0)p_0$$

General optimal classification rule

- The misclassification probability or, in general, the expected cost of misclassification can be used to compare different classification rules.
- We also have the following **general mathematical result**:
  ECM is minimized by choosing

$$R = \begin{cases} 0; & \dfrac{P(Y=0|\mathbf{x})}{P(Y=1|\mathbf{x})} > \dfrac{c(0|1)}{c(1|0)} \\ 1; & \dfrac{P(Y=1|\mathbf{x})}{P(Y=0|\mathbf{x})} > \dfrac{c(1|0)}{c(0|1)} \end{cases}$$

- This shows that if there is no misclassification costs, then the rule that minimizes misclassification probability is given by

$$R = \begin{cases} 0; & \dfrac{P(Y=0|\mathbf{x})}{P(Y=1|\mathbf{x})} > 1 \\ 1; & \dfrac{P(Y=1|\mathbf{x})}{P(Y=0|\mathbf{x})} > 1 \end{cases}$$

Probability ratio rule

- The optimality is shown in Assignment 4, i.e. it is shown that the following rule

$$R = \begin{cases} 0; & \dfrac{P(Y=0|\mathbf{x})}{P(Y=1|\mathbf{x})} > 1 \\ 1; & \dfrac{P(Y=1|\mathbf{x})}{P(Y=0|\mathbf{x})} > 1 \end{cases}$$

  has the smallest chance of misclassification.
- We observe that the rule is based on the probability ratio.
- The probability ratio has a natural interpretation:

Choose what is more probable!

- Since the log is an increasing function, one can use the log-likelihood ratio (**and no!, the log of the ratio is not the ratio of logs**):

$$R = \begin{cases} 0; & \dfrac{\log P(Y=0|\mathbf{x})}{\log P(Y=1|\mathbf{x})} > 1 \\ 1; & \dfrac{\log P(Y=1|\mathbf{x})}{\log P(Y=0|\mathbf{x})} > 1 \end{cases}$$

Posterior probability ratio vs. likelihood ratio

- Given features $\mathbf{x}_0$, the posteriori probabilities are $P(Y=0|\mathbf{x}_0)$ and $P(Y=1|\mathbf{x}_0)$.
- These do not require prior for Y neither the assumption of randomness of $\mathbf{X}$.
- Define

$$R(\mathbf{x}_0) = \begin{cases} 0; & P(Y=0|\mathbf{x}_0) > P(Y=1|\mathbf{x}_0) \\ 1; & \text{otherwise} \end{cases}$$

- If X is random and the prior distribution of Y is given, then

$$\frac{P(Y=0|\mathbf{x})}{P(Y=1|\mathbf{x})} = \frac{P(\mathbf{x}|Y=0)P(Y=0)}{P(\mathbf{x}|Y=1)P(Y=1)} = \frac{f_0(\mathbf{x})p_0}{f_1(\mathbf{x})p_1}$$

- If $p_0 = p_1$, then the classification is equivalent to the one that is based on the **fitted likelihood ratio** of $\mathbf{X}$.

Two normal populations different in means

- Suppose $f_i(\mathbf{x})$ is $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 0, 1$.

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i))$$

so that

$$\ln\left(\frac{f_0(\mathbf{x})}{f_1(\mathbf{x})}\right) = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)/2$$

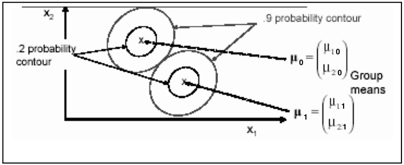
Linear classification rule: Take $R = 0$ if

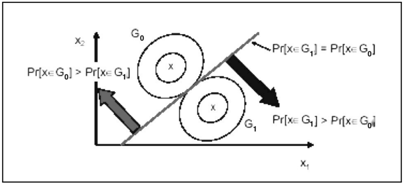
$$(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1)$$

$$\geq \ln(p_1/p_0)$$

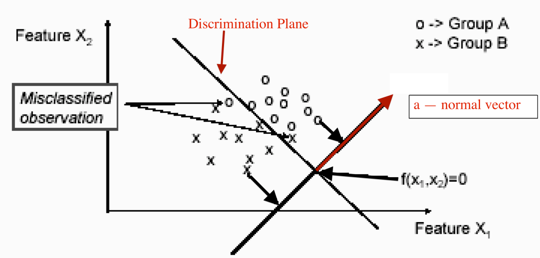
Linear classification – likelihood for the normal case

- The following graphs illustrate the method when there are just two features used to classify





- Thus it corresponds to the geometric rule we mentioned without reference to probability distributions



Discrimination Step – Estimating from the data

For unknown $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$ these are estimated by $\bar{\mathbf{x}}_i, i = 0, 1$ and

$$S = \frac{(n_1 - 1)S_1 + (n_0 - 1)S_0}{n_1 + n_0 - 2}$$

With $y = (\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)' S^{-1} \mathbf{x} = \hat{\ell}' \mathbf{x}$ and

$$y_i = (\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)' S^{-1} \bar{\mathbf{x}}_i = \hat{\ell}' \bar{\mathbf{x}}_i$$

Some simple algebra leads to the classification rule.
Classification rule: Classify $\mathbf{x}$ into $\mathbf{G}_0$ ($Y = 0$) if

$$y > \frac{1}{2}(y_0 + y_1)$$

Linear discriminant function

The case $\mathbf{\Sigma}_0 \neq \mathbf{\Sigma}_1$

$$\ln\left(\frac{f_0(\mathbf{x})}{f_1(\mathbf{x})}\right) = -\frac{1}{2}\mathbf{x}'(\mathbf{\Sigma}_0^{-1} - \mathbf{\Sigma}_1^{-1})\mathbf{x} + (\mu_0'\mathbf{\Sigma}_0^{-1} - \mu_1'\mathbf{\Sigma}_1^{-1})\mathbf{x}$$
$$-\frac{1}{2}\ln\left(\frac{|\mathbf{\Sigma}_0|}{|\mathbf{\Sigma}_1|}\right) - \frac{1}{2}(\mu_0'\mathbf{\Sigma}_0^{-1}\mu_0 - \mu_1'\mathbf{\Sigma}_1^{-1}\mu_1)$$

Classification rule is: Classify $\mathbf{x}$ into $\mathbf{G}_0$ ($Y = 0$) if

$$-\frac{1}{2}\mathbf{x}'(\mathbf{\Sigma}_0^{-1} - \mathbf{\Sigma}_1^{-1})\mathbf{x} + (\mu_0\mathbf{\Sigma}_0^{-1} - \mu_1\mathbf{\Sigma}_1^{-1})\mathbf{x}$$
$$\geq k + \ln(p_2/p_1)$$

where

$$k = \frac{1}{2}\ln(|\mathbf{\Sigma}_0|/|\mathbf{\Sigma}_1|) + \frac{1}{2}(\mu_0'\mathbf{\Sigma}_0^{-1}\mu_0 - \mu_1'\mathbf{\Sigma}_1^{-1}\mu_1)$$

Quadratic discriminant function

Classification based on data – testing phase

- Classification rules based on observations give regions $\hat{\mathcal{X}}_0, \hat{\mathcal{X}}_1$.
  AER=Actual Error Rate

$$\text{AER} = p_0 \int_{\hat{\mathcal{X}}_1} f_0(\mathbf{x})\, d\mathbf{x} + p_1 \int_{\hat{\mathcal{X}}_0} f_1(\mathbf{x})\, d\mathbf{x}$$

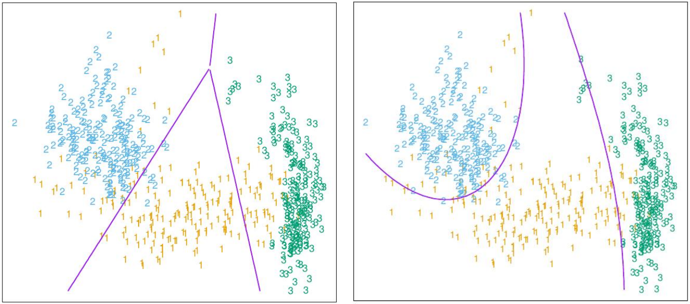
- AER can be estimated by APER (apparent error rate) based on the "confusion matrix":

		Predicted belonging to		
		$\mathbf{G}_0$	$\mathbf{G}_1$	
Actual belonging to	$\mathbf{G}_0$	n_{0c}	n_{0m}	n_0
	$\mathbf{G}_1$	n_{1m}	n_{1c}	n_1

- APER=Apparent Error Rate=$\frac{n_{0m}+n_{1m}}{n_0+n_1}$=the proportion misclassified.

Illustration of linear and quadratic classifications

- Methods extend to more than just two groups
- Here we illustrate the linear and quadratic classification into three classes



- One can use **(cross)validation step** to chose between the two methods of classification

Mixture model

- We assume that the feature data $\mathbf{X}$'s are coming from two different models.
- The two models are possible and from which model the data are arriving is indicated by a binary (generally unobserved) variable Y

$$
\begin{aligned}
\mathbf{X}_0 &\sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0^2) \\
\mathbf{X}_1 &\sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1^2) \\
\mathbf{X} &= (1 - Y)\mathbf{X}_0 + Y\mathbf{X}_1,
\end{aligned}
$$

- We assume that Y is equal 0 or 1, with probabilities p_0 and $p_1 = 1 - p_0$, respectively.

Complete model for the data

- Density

$$g_{\mathbf{X},Y}(\mathbf{X},Y) = \begin{cases} p_0\phi_{\theta_0}(\mathbf{x}) & : Y = 0 \\ p_1\phi_{\theta_1}(\mathbf{x}) & : Y = 1. \end{cases}$$

- the densities ϕ_{θ_0}, ϕ_{θ_1} do not need to be normal although we focus on this case, for illustration.
- Parameters: $\theta = (p_0, \theta_0, \theta_1) = (p_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$
- Full data loglikelihood

$$\begin{aligned} l(\theta; \mathbf{x}_i, y_i) = & \sum_{i=1}^{N} \left((1 - y_i) \log\left(\phi_{\theta_0}(\mathbf{x}_i)\right) + y_i \log\left(\phi_{\theta_1}(\mathbf{x}_i)\right)\right) \\ & + \sum_{i=1}^{N} \left((1 - y_i) \log p_0 + y_i \log p_1\right) \end{aligned}$$

Training phase

- We note that for the training data we assume that Y_i's are given.
- The MLE of $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ would be the sample means and sample covariances corresponding values of $\mathbf{x}_i$' and the estimate of p_1 would be the proportion of Y_i's that are equal to one.
- In the general case of an arbitrary distribution ϕ_θ we find the MLE of θ (or any other suitable method) by whatever means that are available for this distribution.

Classification rule

- Classification can based on

$$R = \begin{cases} 0; & \dfrac{P(Y=0|\mathbf{x})}{P(Y=1|\mathbf{x})} = \dfrac{\phi_{\hat{\theta}_0}(\mathbf{x})\hat{p}_0}{\phi_{\hat{\theta}_1}(\mathbf{x})\hat{p}_1} > \dfrac{c(0|1)}{c(1|0)} \\[2ex] 1; & \dfrac{P(Y=1|\mathbf{x})}{P(Y=0|\mathbf{x})} = \dfrac{\phi_{\hat{\theta}_1}(\mathbf{x})\hat{p}_1}{\phi_{\hat{\theta}_0}(\mathbf{x})\hat{p}_0} > \dfrac{c(0|1)}{c(1|0)} \end{cases}$$

Final remarks

- We have seen several different approaches to the classification problem.
- It is not obvious a'priori which one will work for a given data set.
- **Step One:** This is the nature of the data mining approach to try several such methods on the training data
- **Step Two:** Validate the best one based on validation
- **Step Three:** Test the chosen one on the test data.
- Only then, one should propose it for the use outside of available data sets
- The methods could be sequentially improved once the new data for classification are arriving

Quotation

The classification of facts, the recognition of their sequence and relative significance is the function of science, and the habit of forming a judgment upon these facts unbiased by personal feeling is characteristic of what may be termed the scientific frame of mind.

Karl Pearson *The Grammar of Science* **(1900)***

By Elliott & Fry - N.P.G.

*The founder of the world's first university statistics department at University College London