

# Machine Learning and Programming in Python

Lecture for Master and PhD students

Chair of Data Science in Economics

Ruhr University Bochum

Summer semester 2024

Lecture 6

## Cross-validation

- Cross-validation is a resampling method
- Draw samples from the training data set, fit model, investigate how results differ for different samples
- For the methods of regularisation:  
choose  $\lambda$  via cross-validation

- k-fold cross-validation:

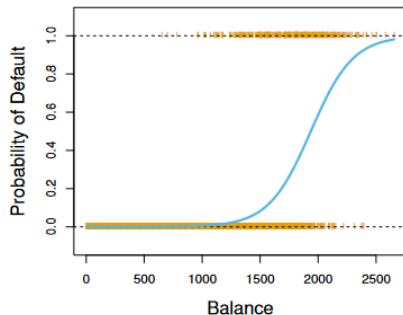
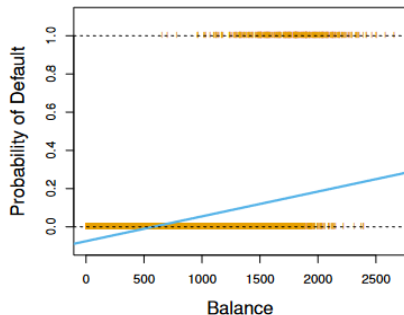
- ▶ Randomly split the training data into  $k$  groups (folds) of roughly equal size
- ▶ Pick a grid of  $\lambda$  values, and for each successive value:
  - ★ Set aside the first fold as a validation set
  - ★ Fit the model on the remaining folds,  $i=2,\dots,k$
  - ★ Compute the  $MSE_1$  on the validation set
  - ★ Repeat by using the second fold as the validation set and fitting on folds  $i=1,3,\dots,k$
  - ★ Compute the cross-validation error,  $CV_k = \frac{1}{k} \sum_{i=1}^k MSE_i$
- ▶ Select the  $\lambda$  which gives the lowest cross-validation error

## Classification

- Classification methods are also an approach to **Supervised Learning**.
- Difference to regression problems: consider here a qualitative variable (instead of a quantitative variable for regression problems)

- Qualitative variables take values in an unordered set  $C$ , such as:  
eye color: *brown, blue, green*  
email: *spam, nospam*
- Given a feature vector  $X$  and a qualitative response  $Y$  taking values in the set  $C$ , the classification task is to build a function  $C(X)$  that takes as input the feature vector  $X$  and predicts its value for  $Y$
- Often we are more interested in estimating the probabilities that  $X$  belongs to each category in  $C$
- For example, it is more valuable to have an estimate of the probability that an insurance claim is fraudulent, than a classification fraudulent or not

- Suppose for the Default classification task that we code:  
 $Y = 0$  if No and  $Y = 1$  if Yes
- Can we simply perform a linear regression of  $Y$  on  $X$  and classify as Yes if  $\hat{Y} > 0.5$ ?
  - ▶ In this case of a binary outcome, linear regression does a good job as a classifier
  - ▶ Since in the population  $E(Y|X = x) = P(Y = 1|X = x)$ , we might think that regression is perfect for this task.
  - ▶ However, linear regression might produce probabilities less than zero or bigger than one. Logistic regression is more appropriate.
- If we have a response variable with more than two possible values, and coding suggests an ordering, then linear regression is not appropriate  $\Rightarrow$  Multiclass Logistic Regression or Discriminant Analysis are more appropriate



The orange data points indicate the response  $Y$ , either 0 or 1.  
Linear regression does not estimate  $P(Y = 1|X)$  well.  
Logistic regression seems well suited to the task.

## Logistic Regression

- Logistic regression uses the form:

$$p(X) = P(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

- $p(X)$  will have values between 0 and 1
- Log odds:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$



- We use maximum likelihood to estimate the parameters.
- $l(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i))$
- This likelihood gives the probability of the observed zeros and ones in the data. We pick  $\beta_0$  and  $\beta_1$  to maximize the likelihood of the observed data.
- We can compute the estimated probability  $\hat{p}(X)$  at some  $X = x$  with the  $\hat{\beta}$  coefficient estimates

- Having more than two classes: multi-class logistic regression = multinomial regression

$$\Pr(Y = k|X) = \frac{e^{\beta_{0k} + \beta_{1k}X_1 + \dots + \beta_{pk}X_p}}{\sum_{\ell=1}^K e^{\beta_{0\ell} + \beta_{1\ell}X_1 + \dots + \beta_{p\ell}X_p}}$$

## Discriminant Analysis

- Here the approach is to model the distribution of  $X$  in each of the classes separately, and then use Bayes theorem to flip things around and obtain  $P(Y|X)$ .
- When we use normal (Gaussian) distributions for each class, this leads to linear or quadratic discriminant analysis.
- However, this approach is quite general, and other distributions can be used as well. We will focus on normal distributions

- Bayes theorem (for classification):

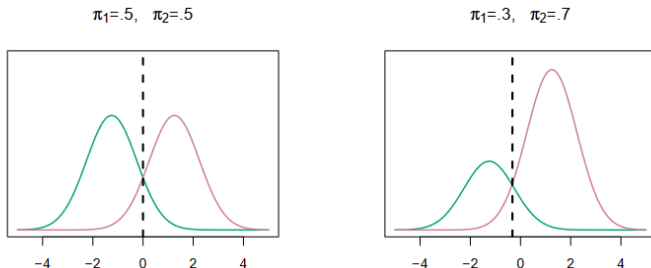
$$P(Y = k|X = x) = \frac{P(X=x|Y=k)P(Y=k)}{P(X=x)}$$

- One writes this slightly differently for discriminant analysis:

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^k \pi_l f_l(x)}, \text{ where}$$

$f_k(x) = P(X = x|Y = k)$  is the density for  $X$  in class  $k$ . Here we will use normal densities for these, separately in each class

$\pi_k = P(Y = k)$  is the marginal or prior probability for class  $k$



- We classify a new point according to which density is highest.
- When the priors are different, we take them into account as well, and compare  $\pi_k f_k(x)$ . On the right, we favor the pink class - the decision boundary has shifted to the left

## Why discriminant analysis?

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis does not suffer from this problem.
- If  $n$  is small and the distribution of the predictors  $X$  is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model.
- Linear discriminant analysis is popular when we have more than two response classes, because it also provides low-dimensional views of the data

## Linear Discriminant Analysis when $p = 1$

The Gaussian density has the form:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

Here  $\mu_k$  is the mean, and  $\sigma_k^2$  the variance (in class  $k$ ). We will assume that all the  $\sigma_k^2 = \sigma$  are the same.

Plugging this into Bayes formula, we get a rather complex expression:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_{l=1}^K \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}}$$

## Discriminant functions

To classify at the value  $X = x$ , we need to see which of the  $p_k(x)$  is largest. Taking logs, and discarding terms that do not depend on  $k$ , we see that this is equivalent to assigning  $x$  to the class with the largest discriminant score:

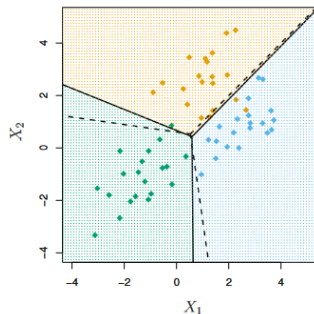
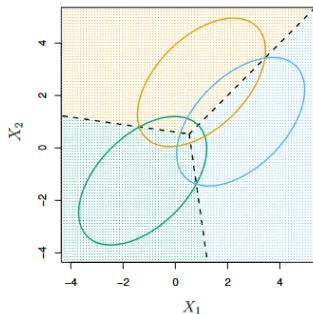
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

Note that  $\delta_k(x)$  is a linear function of  $x$ .

If there are  $K = 2$  classes and  $\pi_1 = \pi_2 = 0.5$ , then one can see that the decision boundary is at

$$x = \frac{\mu_1 + \mu_2}{2}.$$





Here  $p=2$ ,  $K=3$ ,  $\pi_1 = \pi_2 = \pi_3 = 1/3$ . The dashed lines are known as the Bayes decision boundaries. Were they known, they would yield the fewest misclassification errors, among all possible classifiers

- Once we have estimates  $\hat{\delta}_k(x)$ , we can turn these into estimates for class probabilities:

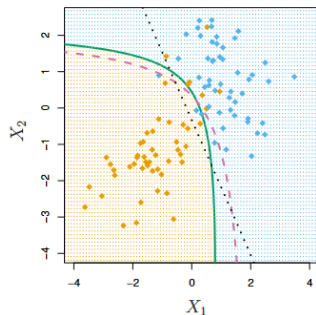
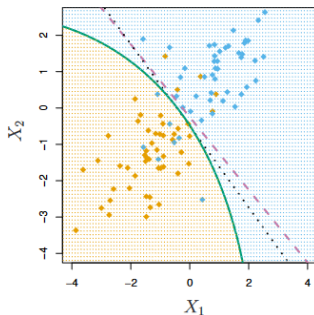
$$\hat{P}(Y = k|X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^K e^{\hat{\delta}_l(x)}}$$

- So classifying to the largest  $\hat{\delta}_k(x)$  amounts to classifying to the class for which  $\hat{P}(Y = k|X = x)$  is largest.
- When  $K = 2$ , we classify to class 2 if  $\hat{P}(Y = 2|X = x) \geq 0.5$ , else to class 1

## Other forms of Discriminant Analysis

- $P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}$
- When  $f_k(x)$  are Gaussian densities, with the same covariance matrix  $\Sigma$  in each class, this leads to linear discriminant analysis. By altering the forms for  $f_k(x)$ , we get different classifiers.
  - ▶ With Gaussians but different  $\Sigma_k$  in each class, we get quadratic discriminant analysis.
  - ▶ With  $f_k(x) = \prod_{j=1}^p f_{jk}(x_j)$  (conditional independence model) in each class we get naive Bayes. For Gaussian this means the  $\Sigma_k$  are diagonal.
  - ▶ Many other forms, by proposing specific density models for  $f_k(x)$ , including nonparametric approaches

# Quadratic Discriminant Analysis



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

Because the  $\Sigma_k$  are different, the quadratic terms matter

## Summary

- Logistic regression is very popular for classification, especially when  $K = 2$ .
- LDA is useful when  $n$  is small, or the classes are well separated, and Gaussian assumptions are reasonable. Also when  $K > 2$ .
- Many other classification methods available, e.g.
  - ▶ K-nearest neighbour (KNN): identify  $K$  points in training data, which lie closest to a point  $x_0$ , then assign  $x_0$  to the class, which has the highest probability given the  $K$  points
  - ▶ Generalised Linear Models (GLM): Output variable is ordinal (count), methods: Poisson Regression, Gamma Regression, Negativ-Binomial Regression

**Literature:**

James, Witten, Hastie, Tibshirani, Taylor (2023), An Introduction to Statistical Learning, Springer, Chapter 4, pp. 135-161; Chapter 5, pp. 201-211.

Hastie, Tibshirani, Friedman (2017), The Elements of Statistical Learning, Springer, Chapter 4, pp. 106-111, pp. 119-122.