Classification, KNN, Logistic Regression (intro)

Dr. Mohamed Elshenawy
mmelshenawy@gmail.com

1

Previous session …

- What is Learning?
- Regression - OLS
- Overfitting and underfitting
- Feature Selection
- Model Selection

2

This session…

- Parametric and non-parametric models
- KNN
- Classification
- Decision boundary
- Logistic Regression (Intro)

3

The session uses content from

- Book: An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0.

4

Parametric Models

- E.g. Linear Regression
- Parametric methods involve a two-step model-based approach:
  - First, we make an assumption about the functional form, or shape, of f. For example, one very simple assumption is that f is linear in X:

$$f(X) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_p X_p$$

  - After the model is selected, we use a certain procedure to estimate the parameters $\theta_0, \theta_1, \ldots, \theta_p$. (In linear regression, we estimated these parameters such that a pre-defined loss function J is minimized)

5

Parametric Models - Advantage

- Assuming a parametric form for f simplifies the problem. It is generally easier to estimate a bounded set of parameters, than it is to fit an entirely arbitrary function f.

6

Parametric Models - Disadvantage

- *The potential disadvantage* of a parametric approach is that the model we choose will usually not match the true unknown form of f.
- If the chosen model is too far from the true f, then our estimate will be poor.

7

Non-parametric Models

- Non-Parametric models are not characterized by a *bounded set* of parameters.
- Non-parametric methods *do not make explicit assumptions about the functional form of f.*
- Instead, they seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly (let the data speak for themselves rather than forcing them to speak through a tiny vector of parameters).

8

Non-parametric Models - Advantage

- These approaches have the potential to accurately fit a wider range of possible shapes for f
- Good when you have sufficient number of examples and you don't want to worry about selecting the right features (no prior knowledge).

9

Non-parametric Models – disadvantage

- Main disadvantage: a very large number of observations (typically far more than is typically needed for a parametric aprpoach) is required in order to obtain an accurate estimate for *f*

10

Parametric and non-parametric models

- **Parametric Models**
  - Advantages:
    - Easier to understand, fast to learn, require much less data
  - Limitations
    - Constrained by a specific form, cannot model complex functional forms, strong assumptions about the data (limited complexity), poor fit if the chosen model is too far from the true function .
  - Popular methods
    - Linear regression
    - Logistic Regression
- **Non-parametric Models**
  - Advantages
    - Flexible: can adapt to a large number of functional forms, no assumptions about the underlying function (fits data better), can result in high performance predictive models.
  - Limitations
    - Requires much more data, overfitting is more likely to occur, harder to explain, slower to learn
  - Popular methods
    - K-nearest neighbors.
    - Decision trees
    - Support vector machines

11

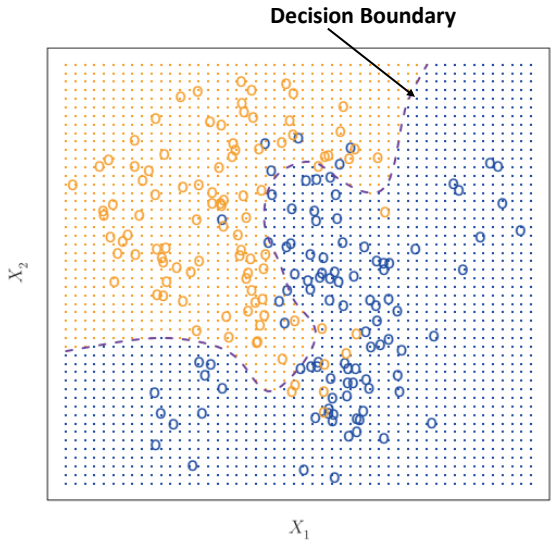
Classification

- We tend to refer to problems with a quantitative response as regression problems, while those involving a qualitative response are often referred to as classification problems
- Classification problems occur often, examples:
  1. An emergency room service that classifies possible medical condition based on a set of symptoms.
  2. An online banking service that classifies if a transaction being is fraudulent based on some user data such as IP address and past transaction history.
  3. An email service that indicates if a received email is spam or not.
  4. Identify which DNA mutations that may cause a give disease and which of these mutations don't based on DNA sequence data for several patients with and without that disease.

An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

12

Classification



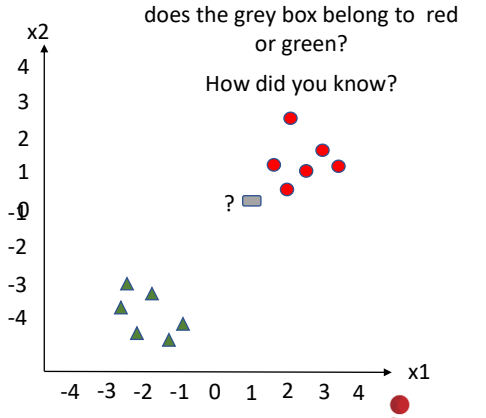
13

KNN

14

K- Nearest Neighbors (KNN) Classifier

- A Non-Parametric Model
- Key idea: the value of the output function is the known output of the nearest training instance.
- An example of *instance-based learning*: it constructs hypotheses directly from the training instances
- A *lazy learning* approach: it delays the use of training set until the system receives a query.



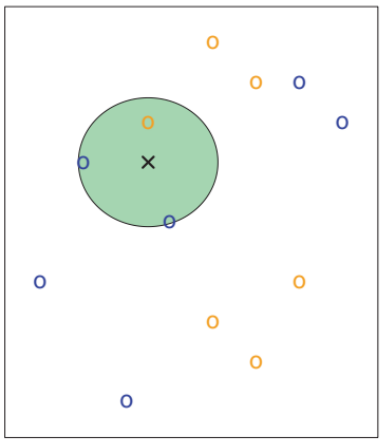
15

How does it work?

- Given a positive integer *K* and a test observation X_t, the KNN classifier first identifies the *K* points in the training data that are closest to X_t represented by $\mathcal{N}_0$.
- It then estimates the conditional probability for class *j* as the fraction of points in $\mathcal{N}_0$ whose response values equal *j*

$$\Pr(Y = j|X = X_t) = \frac{1}{K}\sum_{i\in\mathcal{N}_0} I(y_i = j)$$

$I(y_i = j)$is an *indicator variable* that equals 1 if $y_i = j$



The K-NN using K=3

16

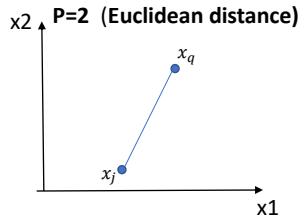
Algorithm

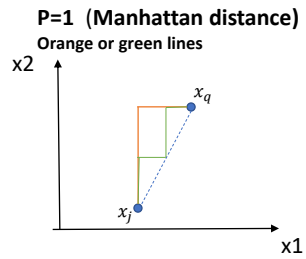
- Algorithm
  - For any given query x_q, find the closest k training instances
  - The output t_q is the class of the majority of k nearest instances

17

How we determine the neighbors? Distances

- Distance from a query point x_q to an example point x_j
- **Minkowski distance** $L^p(x_j, x_q) = (\sum_i (|x_{j,i} - x_q|^p)^{\frac{1}{p}}$
- Use p=2 (Euclidean distance) if the dimensions are measuring similar properties (e.g. width, height, and depth)
- Use p=1 (Manhattan distance) if the dimensions are dissimilar (e.g. age, weight, and gender)

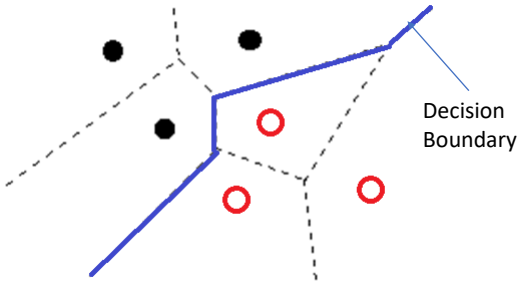




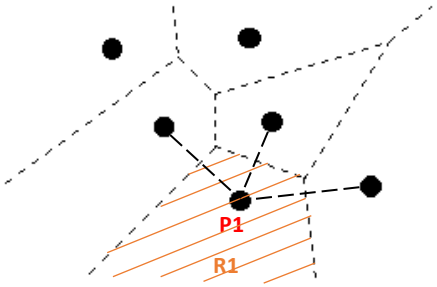
18

Nearest Neighbour

- The value of the output class is calculated
- All points within R1 are closer to P1 than any other point in the training set



cgm.cs.mcgill.ca/



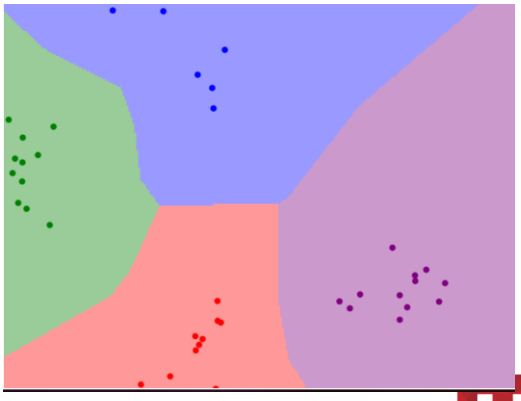
Voronoi diagram

19

KNN

- The value of the output function is calculated by taking the plurality vote of the k nearest neighbors.
- To avoid ties, k is always chosen to be an odd number
- Can we use it for regression? How?

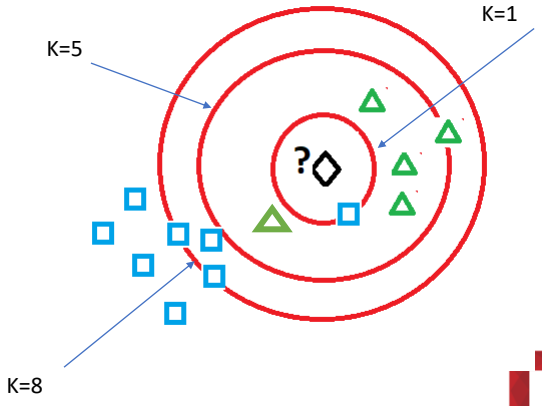
http://vision.stanford.edu/teaching/cs231n-demos/knn/



20

Choosing the right k

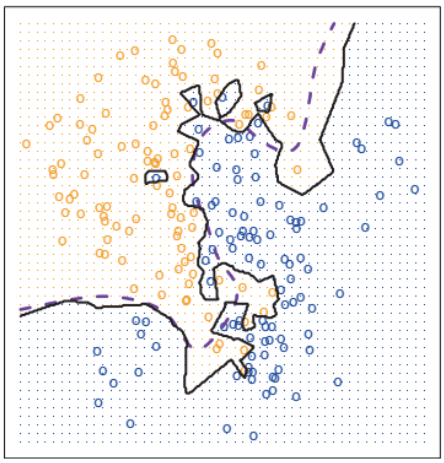
- Too small K means, more sensitive to noise (high variance, low bias).
- Too large K, you will have a less flexible may include instances from other classes (high bias, low variance)



21

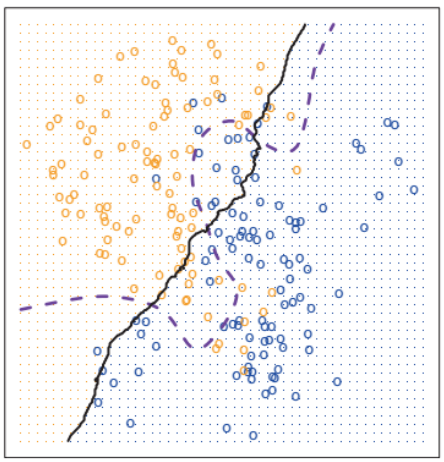
A comparison of the KNN decision boundaries

KNN: K=1



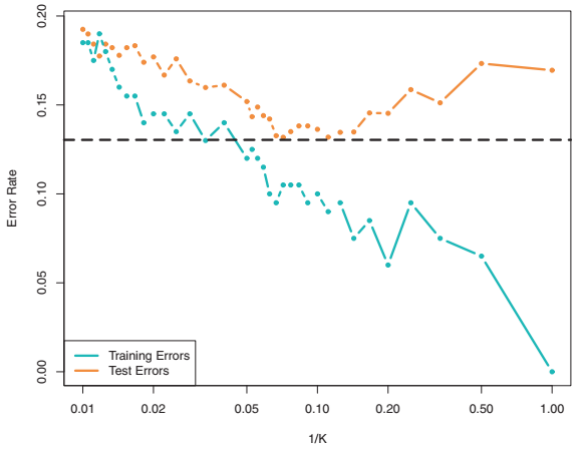
More sensitive to noise

KNN: K=100



22

Test and training errors



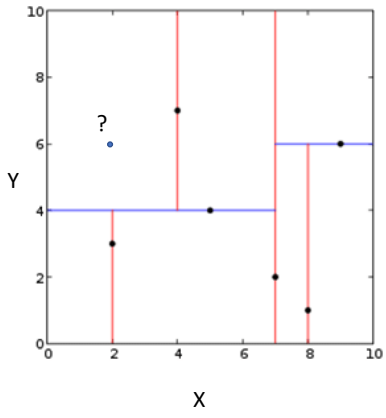
23

Considerations

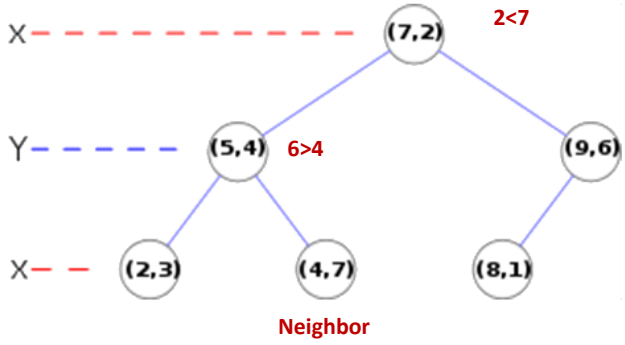
- Nearest Neighbor Search:
  - To find the nearest neighbours we need to find the distance to all N training examples (brute force search) O(dN), d number of dimensions
  - To improve the efficiency and make the algorithm faster we can use data structures such as K-d trees (require pre-sorting) to speed up the search process (via space partitioning mechanism)- good for low-medium dimensions – can miss neighbours – complexity (O(dlogN))
  - Hashing techniques such as Locality sensitive hashing (**LSH**) can be used to approximate the nearest neighbor search in high dimensions.
- Other considerations to improve the result of KNN
  - Curse of dimensionality: required number of training sets increases with the dimension.
  - Consider reducing d by removing irrelevant and correlated dimensions.
  - Consider reducing N by removing redundant data (e.g. condensing)

24

K-d trees

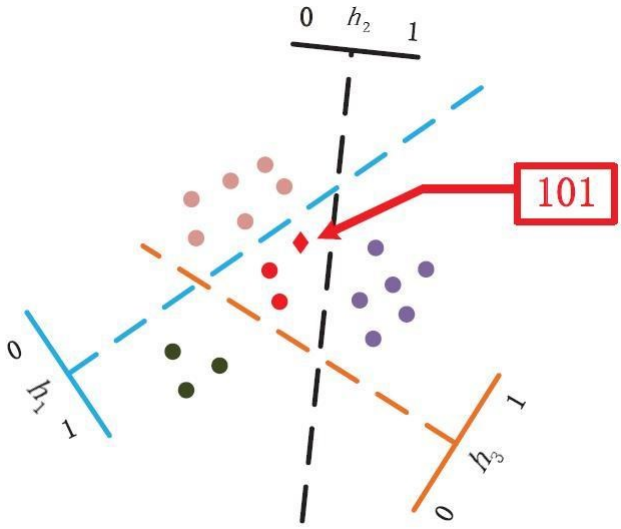


Find neighbor of
(6,2)



25

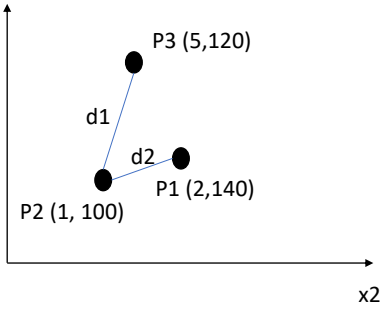
Locality Sensitive Hashing



26

Why we need feature scaling?

- d1 = $sqrt\left((5-1)^2+(120-100)^2\right) = 20.39$
- d2 = $sqrt\left((2-1)^2+(140-100)^2\right) = 40.01$
- Which attribute determine the distance?
- If one attribute has a broad range of values, the distance will be governed by this attribute (will be more important in the classification)
- Objective functions in many algorithms will not work properly without feature scaling.



27

Normalization

- Linearly scale the range (for example: scale the range in [0,1]

$$x_i' = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

- When the value of x_i' is the minimum value in the given dataset, the numerator will be 0, and hence x_i' is 0
- When the value of x_i' is the maximum value in the in the given dataset, the numerator is equal to the denominator and thus the value of x_i' is 1
- If the value of x_i' is between the minimum and the maximum value, then the value of x_i' is between 0 and 1
- Disadvantage: you end up with smaller standard deviations, which can suppress the effect of outliers

28

Standardization

- Another scaling technique that scales the feature values so that the mean of the scaled attribute becomes zero and the standard deviation is 1.
- Scale each dimension to have zero-mean and unit-variance

$$x_i' = \frac{x_i - \bar{x}}{\sigma}$$

- where $\bar{x}$ is the mean of the feature vector and σ is the standard variation

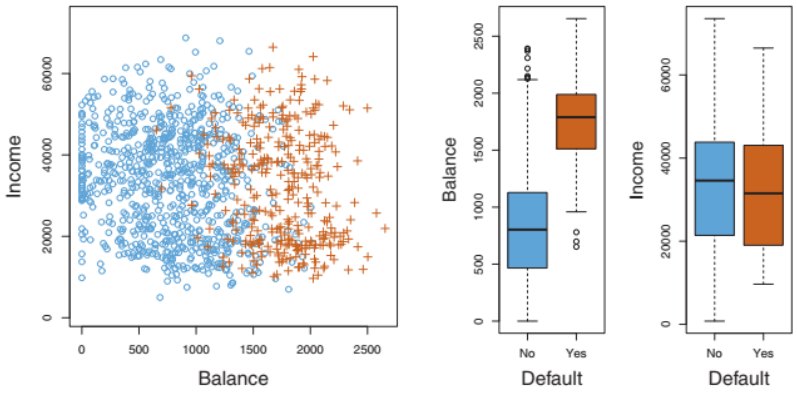
29

Logistic Regression (Intro)

30

Classification Example

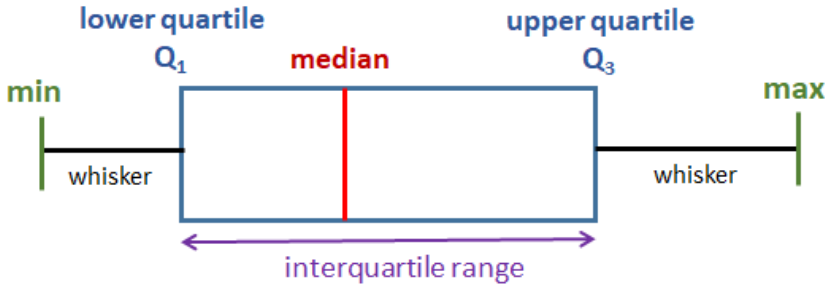
- The individuals who defaulted failed to make a payment on your credit card) on their credit card payments are shown in orange, and those who did not are shown in blue (how can we classify them).
- Individuals who defaulted tended to have higher credit card balances than those who did not.



An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

31

Box and Whisker Plot



https://www.onlinemathlearning.com/box-plot.html

32

Can we use Linear Regression

- Assume you are classifying a credit card transaction and you choose
  - 1 for fraudulent transactions
  - 0 for non-fraudulent transaction
- Can we use linear regression such that it produces y =1 for fraudulent transactions and y=0 for non-fraudulent transaction? What is the problem with this approach?

An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

33

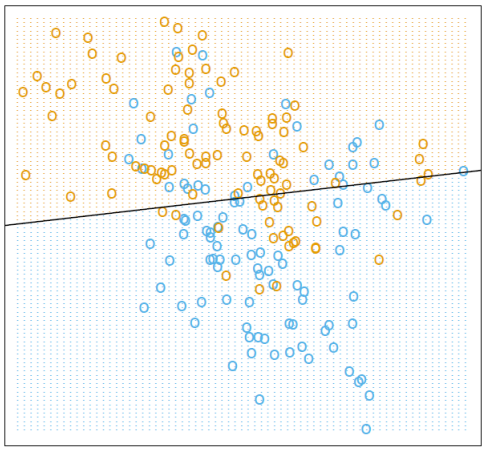
Problem if you try to use linear regression

- One problem is that if you choose a different code (e.g. 0 for fraudulent transactions and 1 for non-fraudulent transaction) will lead to a different model. How about if you use 50 for fraudulent and 150 for non-fraudulent?
- Unless the values of the output response has a natural ordering nature (such as mild, moderate and severe), how can we order different classes (e.g. different medical conditions)

34

Linearly separable problems

- The classifier has a linear decision boundary that separate the space into two regions.
- $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots .$
- Output (o)=
$$\begin{cases} Class\ 1 & if\ \hat{y} \geq 0 \\ Class\ 2 & if\ \hat{y}\ otherwise \end{cases}$$
- o= $sign(\hat{y})$

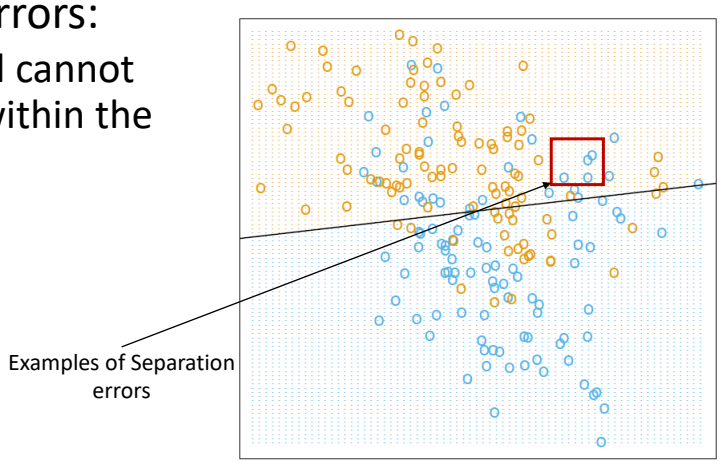


An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

35

How about separation errors

- Causes for separation errors:
  - The model is simple and cannot capture the variations within the data
  - Noise in the data and mislabelling



An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

36

How Can We Estimate the Model Parameters?

- Zero-one loss.

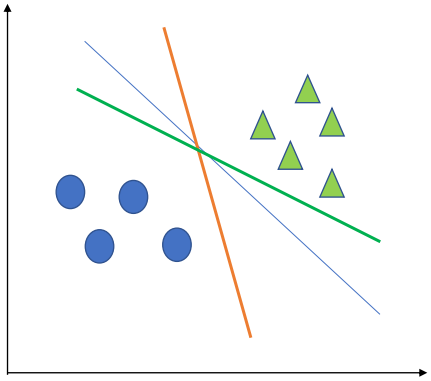
$$l_{0-1}(o,t) = \begin{cases} 1 \; if \; t \neq o \\ 0 \; if \; t = o \end{cases}$$

Is it easy to minimize this function? Why?

- Asymmetric Binary Loss (ABL)

$$l_{ABL}(o,t) = \begin{cases} \alpha \; if \; t = 1 \; and \; o = -1 \\ \beta \; if \; t = -1 \; and \; o = 1 \\ 0 \; if t = o \end{cases}$$

What is the problem with such functions?



37

Logistic Regression

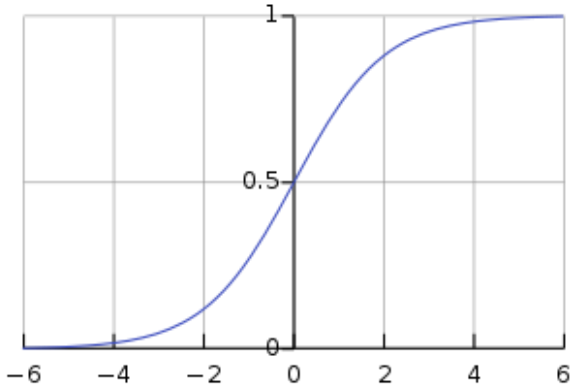
- Logistic regression models the *probability* that the output belongs to a particular category $\Pr(Y = k|X = x)$ using the logistic (sigmoid) function.
- Why a logistic function?
  - It produces a value between 0 and 1 (range of the probability values)
  - It can be viewed as a differentiable and smoothed alternative to sign.
- Allows gradient-based learning of parameters
- A model for classification (not regression).

An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

38

Logistic Function

- $\sigma(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{e^z+1}$

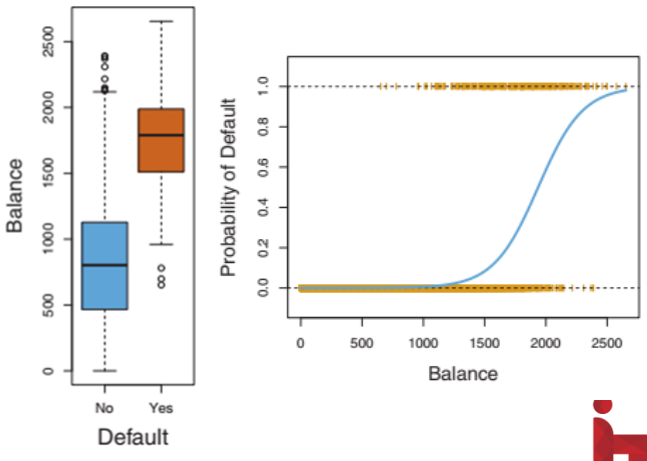


An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

39

In the credit card payments example

- $\Pr(default = Yes|blanace) = \sigma(\theta_0 + \theta_1 * blance)$
- One might predict default = Yes for any individual for whom $\Pr(default = Yes|blanace) > 0.5$.
- Alternatively, a more conservative approach in predicting whether an individual will default (fail to repay) on his or her credit card payment $\Pr(default = Yes|blanace) > 0.1$

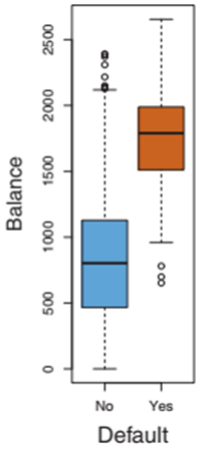


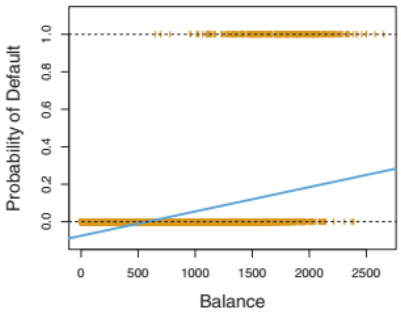
An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

40

Assume you use a linear model (without a logistic function)

- $\Pr(default = Yes|blanace) = \theta_0 + \theta_1 * blance$
- For balances close to zero we predict a negative probability of default; if we were to predict for very large balances, we would get values bigger than 1 (not sensible).





An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

41