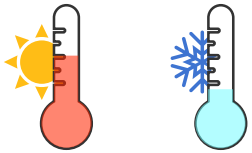
Classification I

Lecture 4

Termeh Shafie

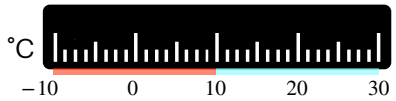
Regression vs. Classification



- What is the temperature going to be tomorrow?
- Will it be hot or cold tomorrow?

continuous variable
(from $-\infty$ to ∞)

categorical variable
(0 to 1)



Generative vs. Discriminative Classification Methods

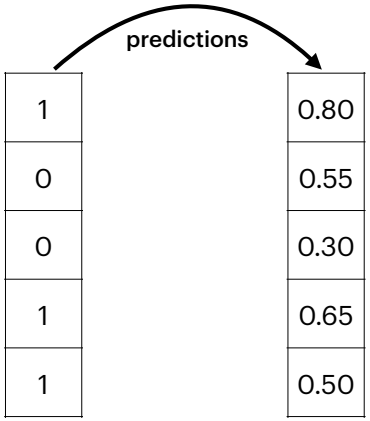
Generative model how data is produced	**Discriminative** model how to separate classes
• Learns a "recipe" for each class • Calculate probability based on recipe of a new point belonging to each class • Example: Naive Bayes, LDA, QDA	• Focus directly on distinguishing classes and not the data generating process • Draw the best possible boundary to separate the classes based on the data • Example: Logistic regression, KNN*, SVM, Tree based methods

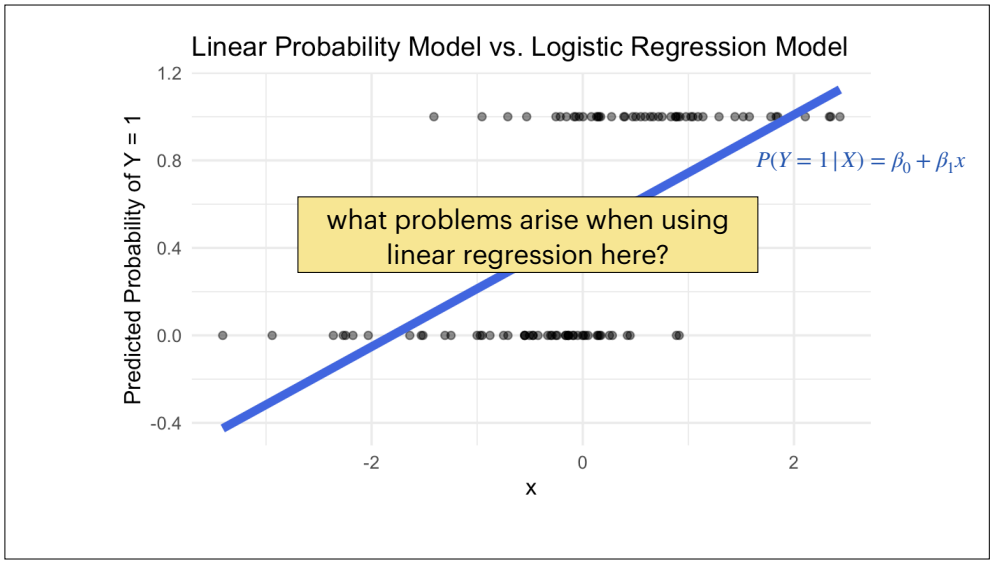
*non-probabilistic but boundary-based

Model $P(Y)$ and $P(X \mid Y)$, derive $P(Y \mid X)$

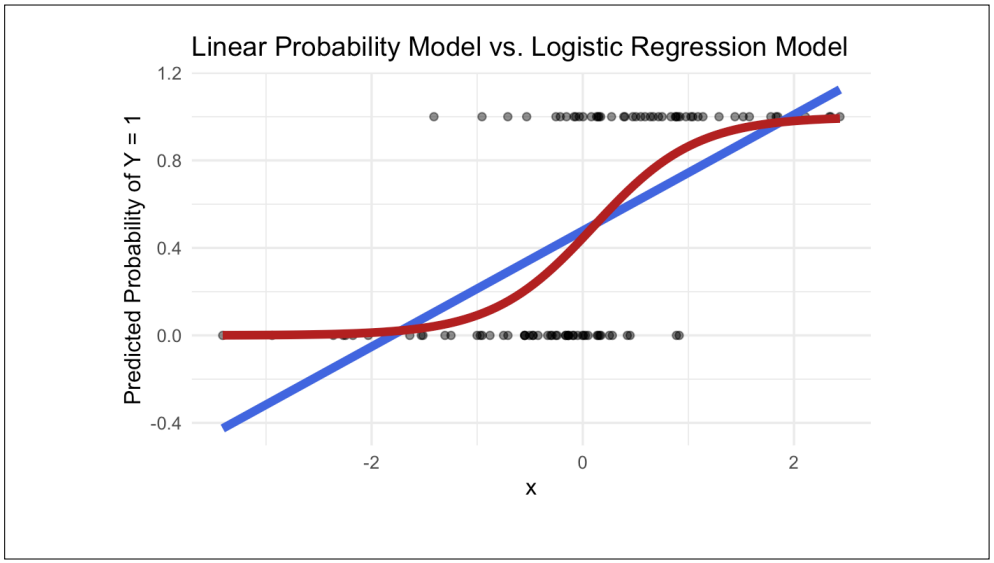
Model $P(Y \mid X)$ directly

Linear Probability Model

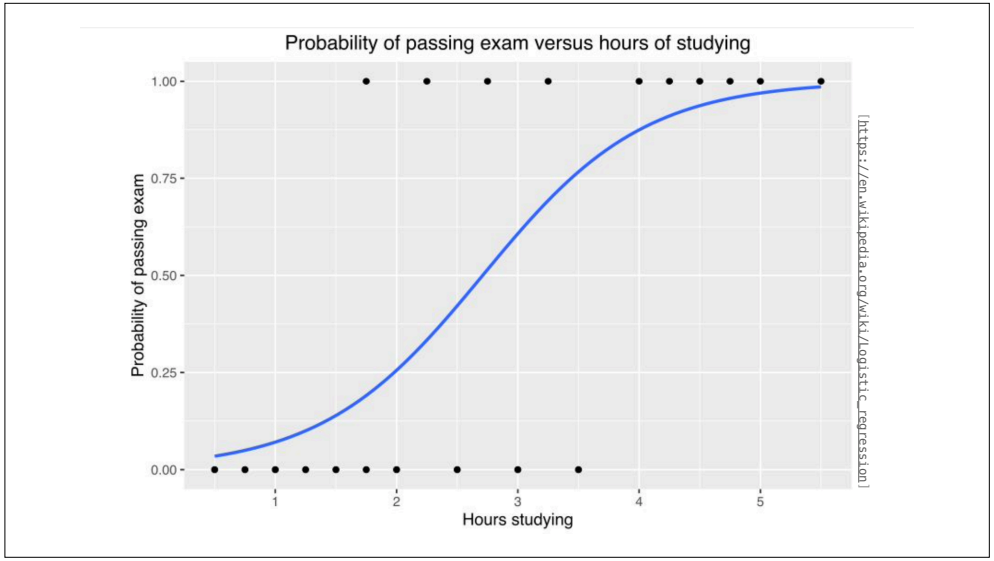




5



6



7

Link Functions

a method to get "non-linear" linear regression
(more on this topic in a later lecture...)

$$y = X\beta$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$y = g^{-1}(X\beta)$$

the link function transforms the probabilities of the levels of a categorical response variable to a continuous scale that is unbounded

the link function transforms back the expectation of the response to the linear function

8

Odds and log Odds

odds are the ratio of something happening to something not happening

the odds of my team winning is 1 to 4: $\frac{1}{4} = \frac{\text{⚽}}{\text{⚽⚽⚽⚽}} = 0.25$

the odds of my team winning is 5 to 3: $\frac{5}{3} = \frac{\text{⚽⚽⚽⚽⚽}}{\text{⚽⚽⚽}} = 1.7$

the probability is the ratio of something happening to **everything that could happen**

$$\frac{\text{⚽}}{\text{⚽⚽⚽⚽⚽}} = \frac{1}{5} = 0.20$$

$$\frac{\text{⚽⚽⚽⚽⚽}}{\text{⚽⚽⚽⚽⚽⚽⚽⚽}} = \frac{5}{8} = 0.625$$

9

Odds and log Odds

the probability is the ratio of something happening to **everything that could happen**

the probability of winning: $\frac{\text{⚽⚽⚽⚽⚽}}{\text{⚽⚽⚽⚽⚽⚽⚽⚽}} = \frac{5}{8} = 0.625$

the probability of losing: $\frac{\text{⚽⚽⚽}}{\text{⚽⚽⚽⚽⚽⚽⚽⚽}} = \frac{3}{8} = 1 - \frac{5}{8} = 0.375$

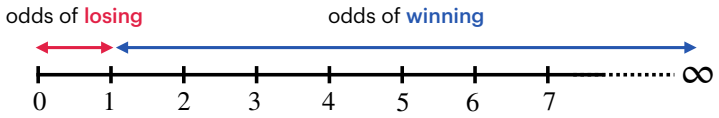
the ratio of probability of winning to the probability of losing $= \frac{p}{(1-p)} = \frac{5/8}{3/8} = \frac{5}{3} = \frac{\text{⚽⚽⚽⚽⚽}}{\text{⚽⚽⚽}}$

the ratio of probabilities are the same as the ratio of the raw counts resulting in the same odds

10

Odds and log Odds

why log odds?



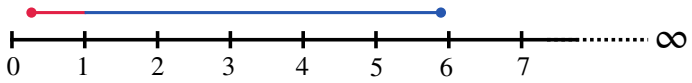
11

Odds and log Odds

why log odds?

example: odds against 1/6=0.17 but odds in favor 6/1=6

taking the log of the odds makes everything symmetrical



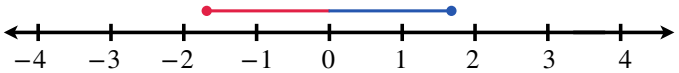
12

Odds and log Odds

why log odds?

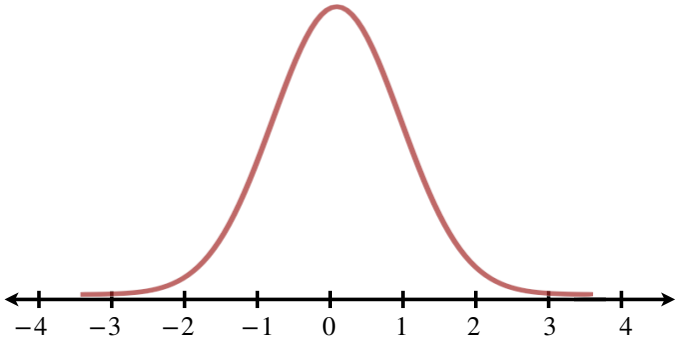
example: log odds against log(0.17)=-1.79 but log odds in favor log(6)=1.79

taking the log of the odds makes everything symmetrical



Odds and log Odds

why log odds?



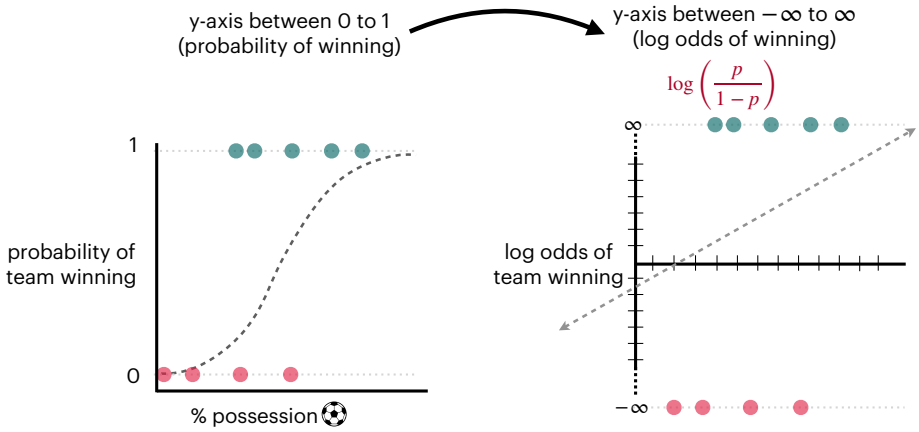
Logit Function

$$y = X\beta$$

$$\log\left(\underbrace{\frac{p}{1-p}}_{\text{odds}}\right) = \beta_0 + \beta_1 x_1$$

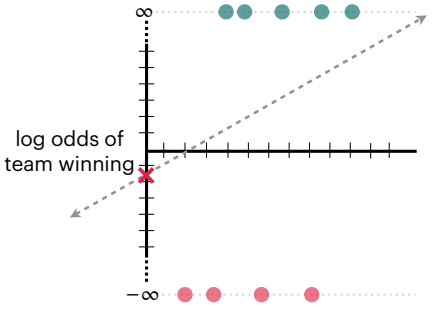
logit link function

Redefining The Response



Model Output

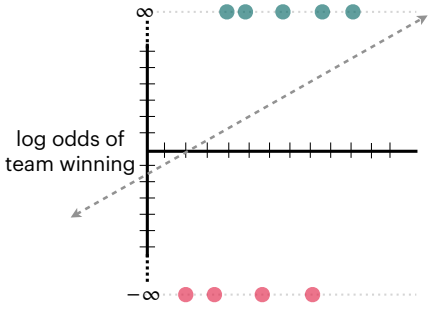
$y = -1.68 + 0.04 \times$ % possession




```
Call:
glm(formula = win ~ possession, family = binomial(), data = df)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.675714   0.414092  -4.047 5.19e-05 ***
possession   0.043467   0.007979   5.447 5.11e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Model Output

$y = -1.68 + 0.04 \times$ % possession



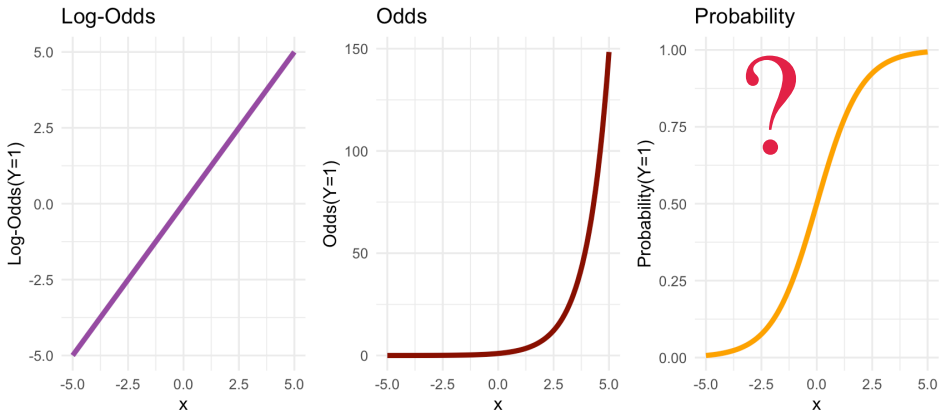

```
Call:
glm(formula = win ~ possession, family = binomial(), data = df)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.675714   0.414092  -4.047 5.19e-05 ***
possession   0.043467   0.007979   5.447 5.11e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

each +1% possession multiplies odds by $\exp(0.04)$

each +10% possession multiplies odds by $\exp(10 \times 0.04)$

⋮

Interpreting Coefficients



Link Functions

$$y = X\beta$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$y = g^{-1}(X\beta)$$

the **link function** transforms the probabilities of the levels of a categorical response variable to a continuous scale that is unbounded

the **link function** transforms back the expectation of the response to the linear function

Link Functions

$$y = X\beta$$

$$\log\left(\underbrace{\frac{p}{1-p}}_{\text{odds}}\right) = \beta_0 + \beta_1 x_1$$

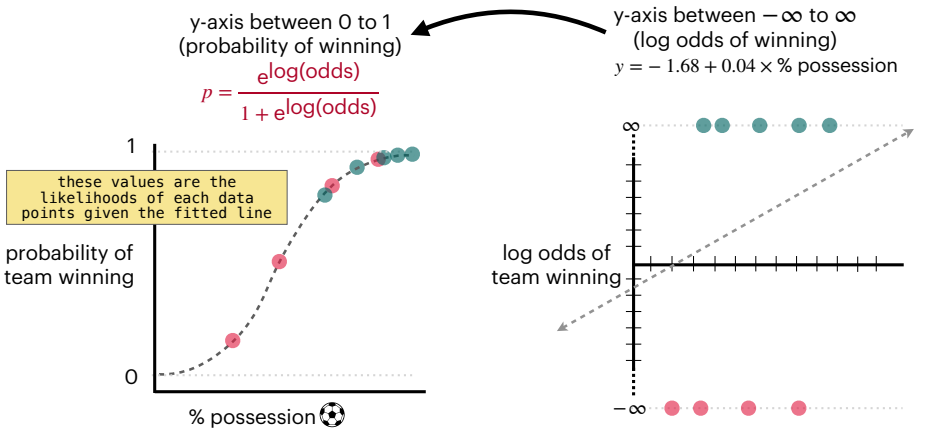
logit link function

$$y = g^{-1}(X\beta)$$

$$p = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}} \quad \text{= [a little algebra]}$$

$$= \frac{1}{1+e^{-(\beta_0+\beta_1 x_1)}}$$

Predicted Probabilities

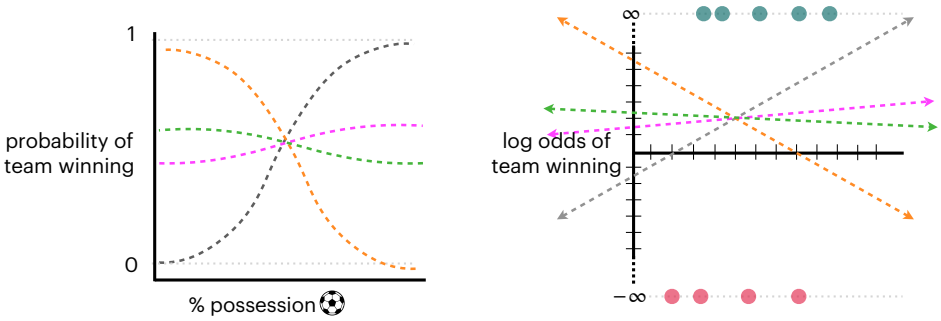


Interpreting Coefficients

Probability p	**Odds** p/(1-p)	**Log Odds** log[p/(1-p)]
0.1	0.1111	-2.1972
0.5	1	0
0.9	9	2.1972

Fitting the Best Line

we fit the line that maximizes the likelihood of the data



Estimating Coefficients: MLE

$$\prod_{i;y_i=1} p(x_i) \qquad \prod_{i;y_i=0} 1 - p(x_i)$$

$$L(\beta_0, \beta_1) = \prod_{i;y_i=1} p(x_i) \cdot \prod_{i;y_i=0} 1 - p(x_i)$$

$$L(\beta_0, \beta_1) = \prod_{i=1} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

$$l(\beta_0, \beta_1) = \sum_{i=1} y_i \log(p(x_i)) + (1 - y_i)\log(1 - p(x_i))$$

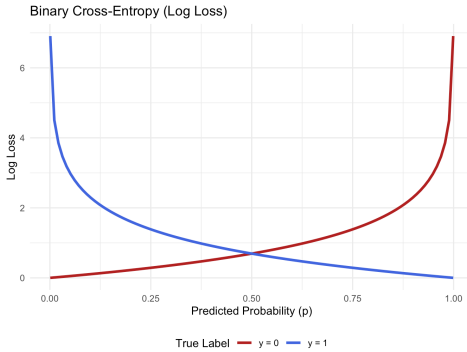
[full proof: https://arunaddagatla.medium.com/maximum-likelihood-estimation-in-logistic-regression-f86ff1627b67

25

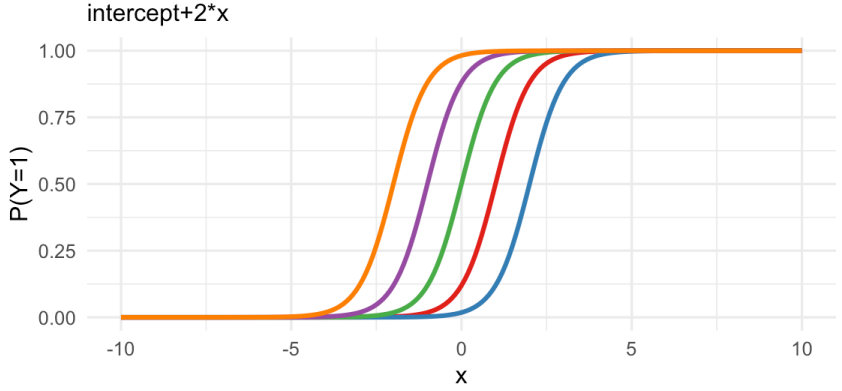
Binary Cross Entropy

$$l(\beta_0, \beta_1) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(p(x_i)) + (1 - y_i)\log(1 - p(x_i))$$

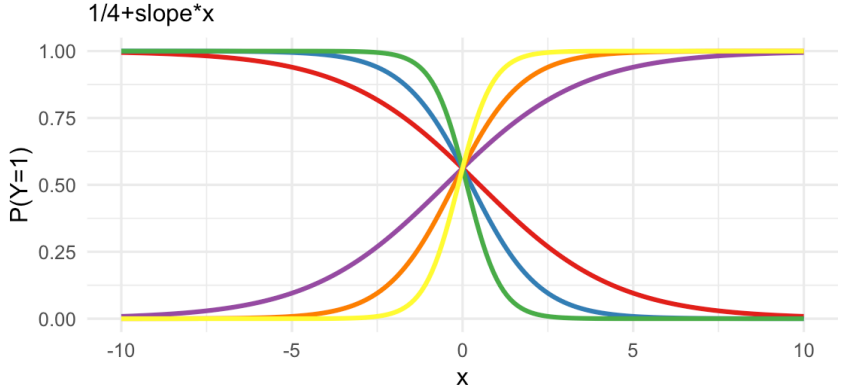
Loss function



26



27



28

Logistic Regression

$$y = X\beta$$

our link function is

$$g(x) = \log \frac{x}{1-x}$$

which has the inverse

$$g^{-1}(x) = \frac{e^x}{1+e^x}$$

$$y = g^{-1}(X\beta)$$ general case

$$p = \frac{e^{X\beta}}{1+e^{X\beta}}$$ specific case

Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\beta = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\beta = \mu$	$\mu = \mathbf{X}\beta$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\beta = -\mu^{-1}$	$\mu = -(\mathbf{X}\beta)^{-1}$
Gamma					
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\beta = \mu^{-2}$	$\mu = (\mathbf{X}\beta)^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\beta = \ln(\mu)$	$\mu = \exp(\mathbf{X}\beta)$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\beta = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\beta)}{1+\exp(\mathbf{X}\beta)} = \frac{1}{1+\exp(-\mathbf{X}\beta)}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\beta = \ln\left(\frac{\mu}{n-\mu}\right)$	
Categorical	integer: $[0, K)$	outcome of single K-way occurrence		$\mathbf{X}\beta = \ln\left(\frac{\mu}{1-\mu}\right)$	
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1				
Multinomial	K-vector of integer: $[0, N]$	count of occurrences of different types $(1, \ldots, K)$ out of N total K-way occurrences			

https://en.wikipedia.org/wiki/Generalized_linear_model#Link_function

Interpreting Logistic Regression Models

- we want to create a spam filter based on 3921 observations/emails
- simple model, one predictor: '`to_multiple`'

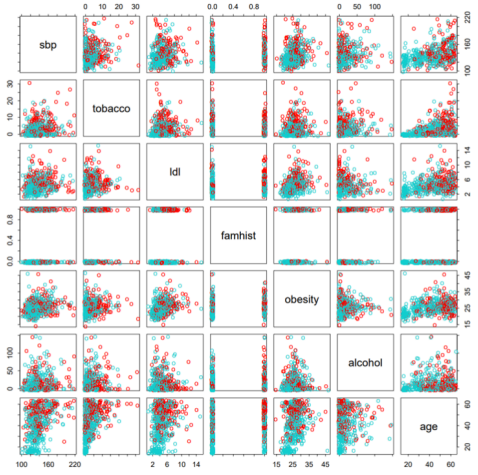
```
Call:
glm(formula = spam ~ to_multiple, family = binomial, data = email)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.477  -0.477  -0.477  -0.477   2.809

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.11609    0.05618 -37.665  < 2e-16 ***
to_multipleyes -1.80918    0.29685  -6.095  1.1e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2437.2  on 3920  degrees of freedom
Residual deviance: 2372.0  on 3919  degrees of freedom
AIC: 2376

Number of Fisher Scoring iterations: 6
```


Example: South African Heart Disease

- From Western Cape,South Africa in early 80s
- Coronary Risk Factor Study (CORIS)
- High incidence of myocardial infarction (MI) in region: 5.1%
- Measurements on seven predictors (risk factors)
- 160 cases, 302 controls. Ages 15-64.
- Outcome is presence/absence of MI at time of survey
- Goal:
  - to identify relative strengths and directions of risk factors
  - intervention study aimed at educating the public on healthier diets

[For more info see ESL 4.4.2]

Example: South African Heart Disease



cases = red
controls = turquoise

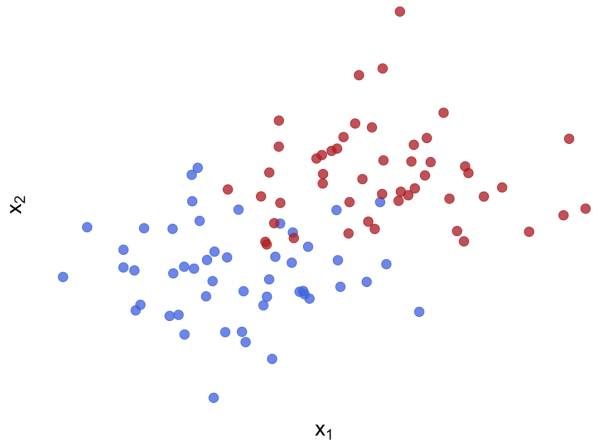
33

Example: South African Heart Disease

term	estimate	std.error	statistic	p.value
(Intercept)	-4.130	0.964	-4.283	0.000
sbp	0.006	0.006	1.023	0.306
tobacco	0.080	0.026	3.034	0.002
ldl	0.185	0.057	3.219	0.001
famhistPresent	0.939	0.225	4.177	0.000
obesity	-0.035	0.029	-1.187	0.235
alcohol	0.001	0.004	0.136	0.892
age	0.043	0.010	4.181	0.000

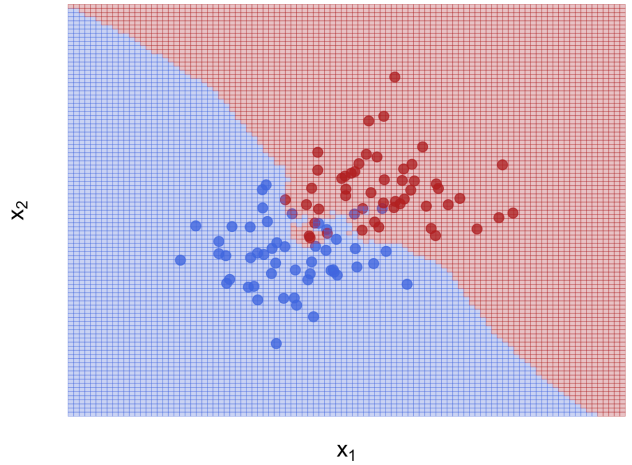
34

K-Nearest Neighbors (KNN)



35

K-Nearest Neighbors (KNN)

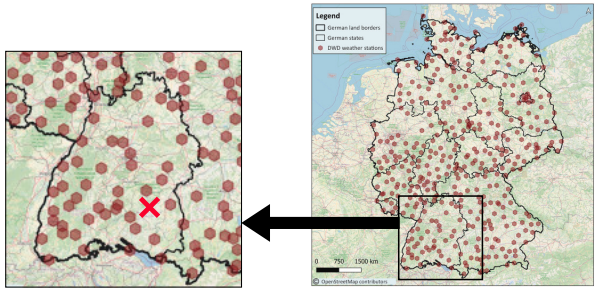


36

K-Nearest Neighbors (KNN)

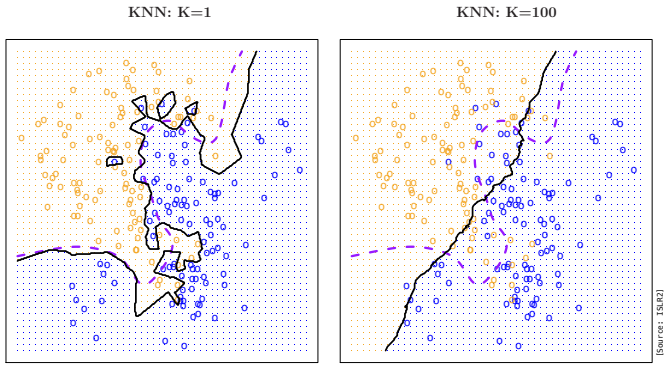
KNN regression tries predicting values of output variable by using a local average

KNN classification attempts predicting the class to which the output variable belong by computing the local probability



The Parameter K

- Smaller K $\Longrightarrow$ Less smooth and more sensitive to noise (more flexible)
- Larger K $\Longrightarrow$ More smooth (less flexible)



The Hyperparameter K

- Smaller K $\Longrightarrow$ Less smooth and more sensitive to noise (more flexible)
- Larger K $\Longrightarrow$ More smooth (less flexible)

how do we choose K?

- choose yourself
- let the data decide (hyperparameter tuning: more on this later...)

The Influence of the Similarity Function

The choice of the distance/similarity function is also important:

- Performance only be good if the distance function encodes "relevant information"

  Example: you want to classify mushrooms as "edible" or "not edible" and as distance function between mushrooms you use the difference in weight...

- Not so obvious sometimes how to define a good distance or similarity function

  Example: you want to classify the genre of songs but how do you compute a similarity between different songs?

This Week's Practical

Logistic Regression and KNN