Classification

DSA 6000: Data Science and Analytics, Fall 2019

Wayne State University

We will learn

- Logistic Regression
  - Model formulation, coefficient interpretation, making predictions
- Assessing model performance, ROC curve
- Linear Discriminant Analysis
- Naïve Bayes Classifier

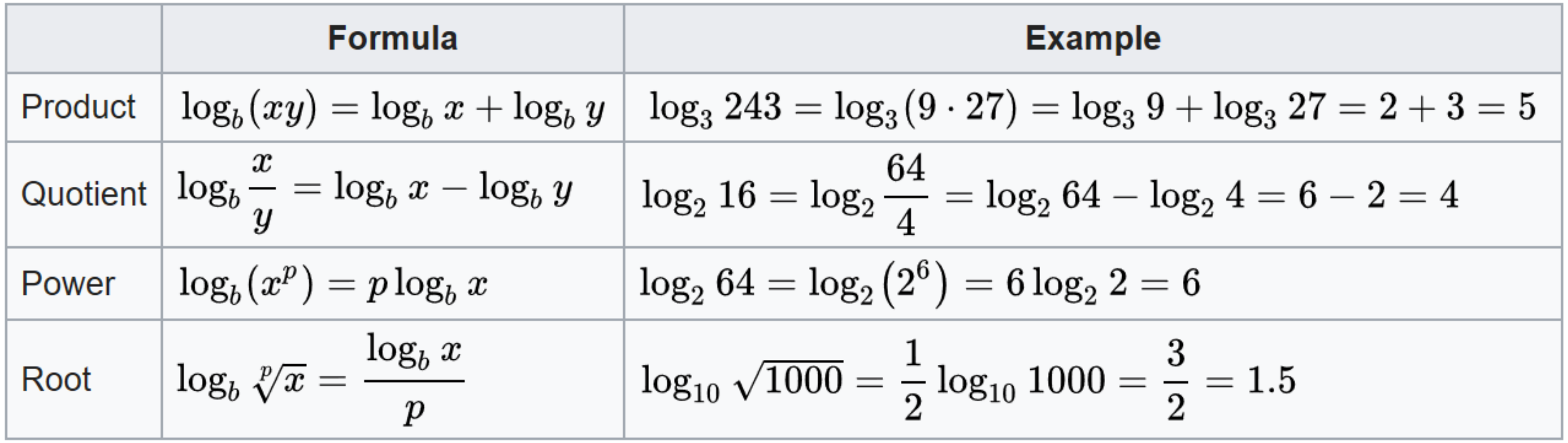
The classification problems

- Response variable is qualitative (or categorical)
  - E.g., Default (yes, no), Reaction to a drug (low, medium, high), Handwritten digits (0, 1, …, 9)
  - A problem with two response levels is a *binary classification* problem
  - A problem with more than two response levels is a *multiclass classification* problem
- Predictors are numeric, $X_1, X_2, \ldots, X_p$
  - For categorical predictors, encode them to numeric using dummy variables
- We will focus on the binary classification problem

Conceptualize the response as an event

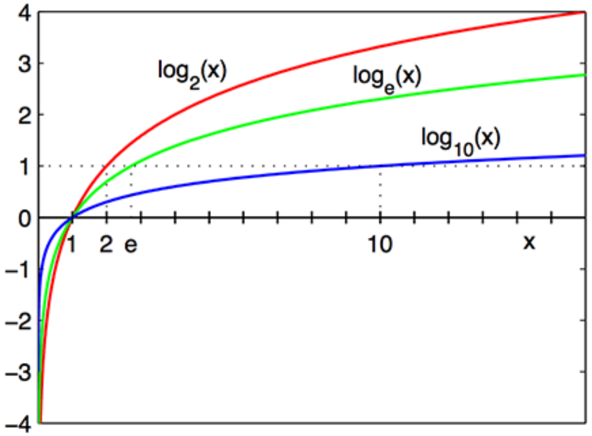
- Describe the meaning of the response variable in a full, definite sentence.
  - This is the most important sentence to convey what your model does
  - It is not only a technical note, but also a business statement
  - A good description can prevent misuse of the model results
- Y = 1 Disconnect the service; Y = 0 Continue with the service
  - What service? Who? What time frame are you talking about? What's the specific action?
- Y = 1 if the customer who has subscribed the Internet + TV bundle service for more than 9 months will unsubscribe this service in the next 2 months; Y = 0 otherwise.
  - It not only describes what is being modeled and predicted, but also implies to the analyst/modeler the scope of the training set, e.g., customers who has been using the service for more than 9 months.

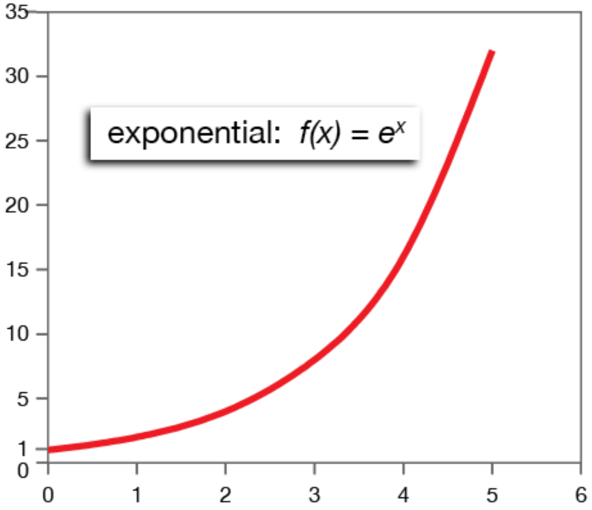
Preliminaries: $\log(x)$ and e^x



	Formula	Example
Product	$\log_b(xy) = \log_b x + \log_b y$	$\log_3 243 = \log_3(9 \cdot 27) = \log_3 9 + \log_3 27 = 2 + 3 = 5$
Quotient	$\log_b \frac{x}{y} = \log_b x - \log_b y$	$\log_2 16 = \log_2 \frac{64}{4} = \log_2 64 - \log_2 4 = 6 - 2 = 4$
Power	$\log_b(x^p) = p \log_b x$	$\log_2 64 = \log_2\left(2^6\right) = 6 \log_2 2 = 6$
Root	$\log_b \sqrt[p]{x} = \frac{\log_b x}{p}$	$\log_{10} \sqrt{1000} = \frac{1}{2} \log_{10} 1000 = \frac{3}{2} = 1.5$

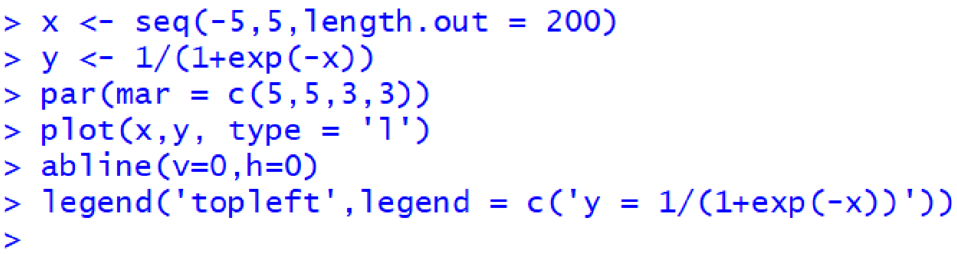
$$e^{x+y} = e^x e^y, \text{ for all } x, y \in \mathbb{R}.$$





Logistic Regression

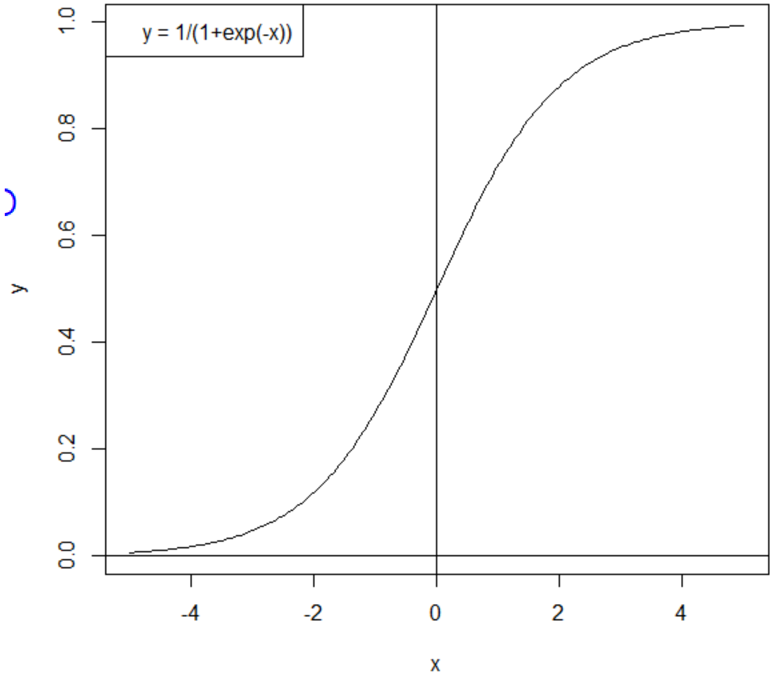
- As in linear regression, we first calculate a linear combination of predictors, $v = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
- Instead of directly using v as the prediction, we transform it using the logit function (also called a sigmoid function, i.e., S-shaped)
  - $\sigma(v) = \frac{1}{1+\exp(-v)}$



```
> x <- seq(-5,5,length.out = 200)
> y <- 1/(1+exp(-x))
> par(mar = c(5,5,3,3))
> plot(x,y, type = 'l')
> abline(v=0,h=0)
> legend('topleft',legend = c('y = 1/(1+exp(-x))'))
>
```

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

$p(X)$ represents the probability of the target **event** given X.



Interpret the coefficients

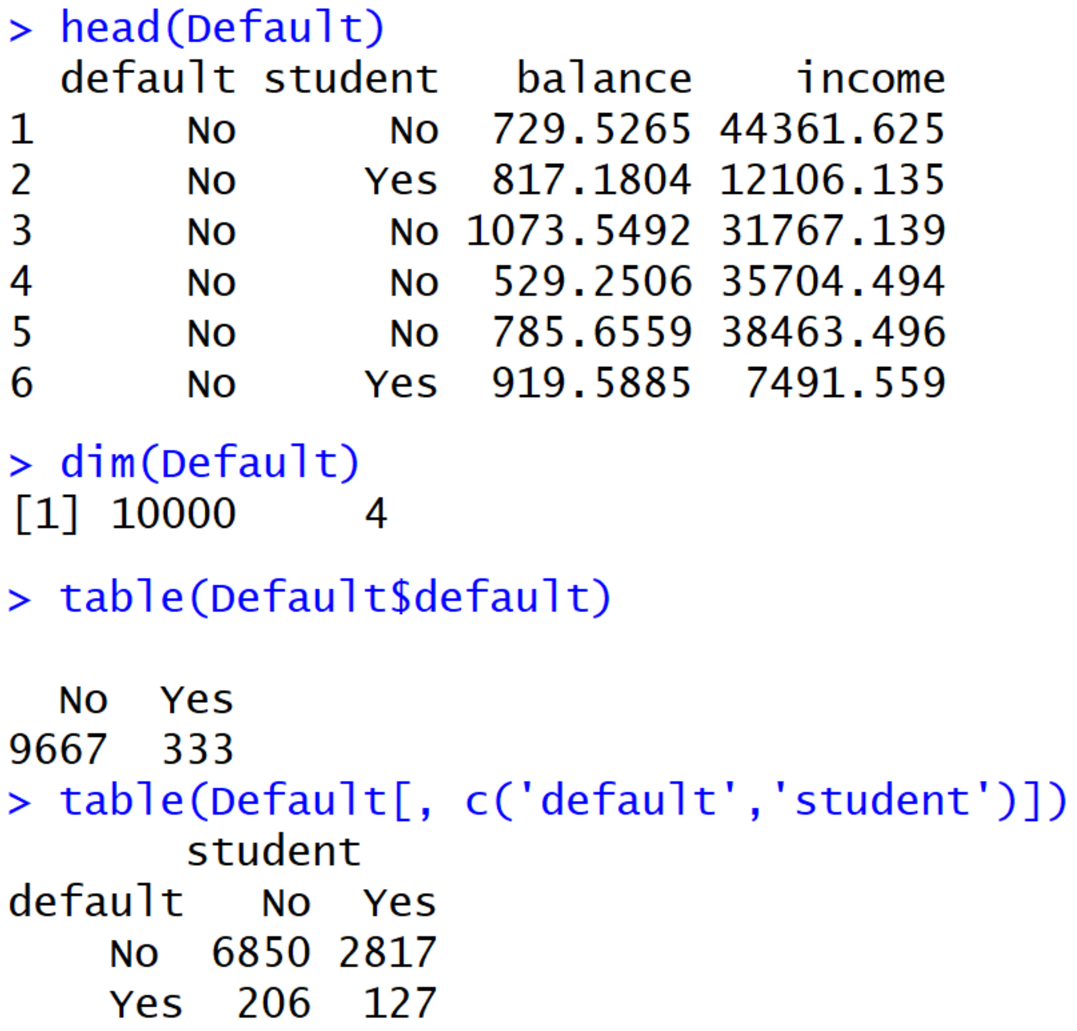
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

- $p(X)$ is the probability that the target event occurs (i.e.,$Y = 1$) given X
- $1 - p(X)$ is the probability that the target event does not occur
- $p(X)/(1 - p(X))$ is called the **odds** of the event
- The log of odds $\log(\frac{p(X)}{1-p(X)})$ is called the **logit** of the event
- Holding other variables unchanged and regardless of their values:
  - β_1 represents the difference in logit incurred by a unit change in X_1
  - $\exp(\beta_1)$ represents the ratio between odds after and before one unit change in X_1

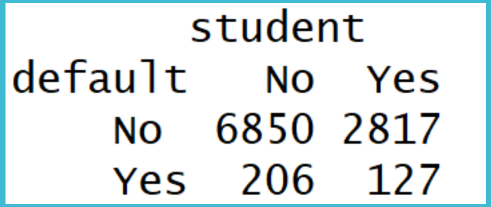
Example

- Let's use the **Default** dataset in the ISLR package.



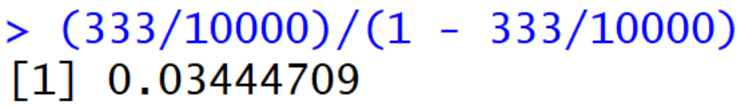
```
> head(Default)
  default student   balance    income
1      No      No  729.5265 44361.625
2      No     Yes  817.1804 12106.135
3      No      No 1073.5492 31767.139
4      No      No  529.2506 35704.494
5      No      No  785.6559 38463.496
6      No     Yes  919.5885  7491.559

> dim(Default)
[1] 10000     4

> table(Default$default)

  No  Yes 
9667  333 
> table(Default[, c('default','student')])
       student
default   No  Yes
    No  6850 2817
    Yes  206  127
```

Working with Odds



```
        student
default   No  Yes
    No  6850 2817
    Yes  206  127
```

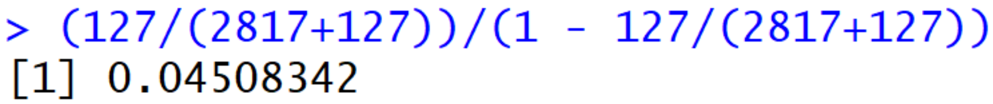
- What is the overall odds of default?



```
> (333/10000)/(1 - 333/10000)
[1] 0.03444709
```

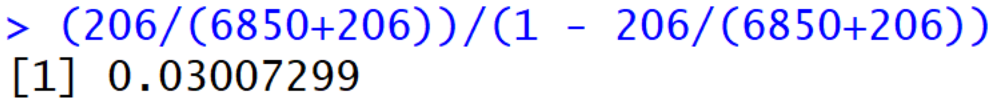
- What is the overall probability of default?
  - 333/10000 = 0.0333

- What is the odds of default among students?



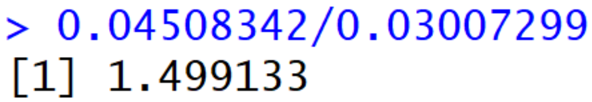
```
> (127/(2817+127))/(1 - 127/(2817+127))
[1] 0.04508342
```

- What is the odds of default among non-students?



```
> (206/(6850+206))/(1 - 206/(6850+206))
[1] 0.03007299
```

- What is the odds ratio between students and non-students?

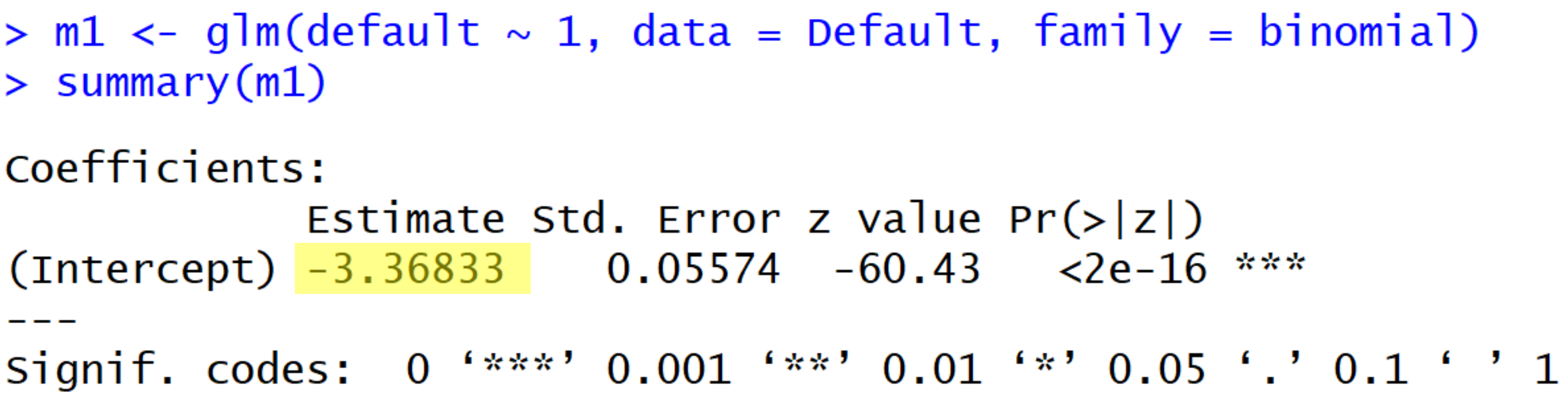


```
> 0.04508342/0.03007299
[1] 1.499133
```

- So the odds of default for students is about 50% higher than the odds of default for non-students

Example

- Let's start with a null model: logistic regression with no predictors
- Model: $\log(\frac{p}{1-p}) = \beta_0$

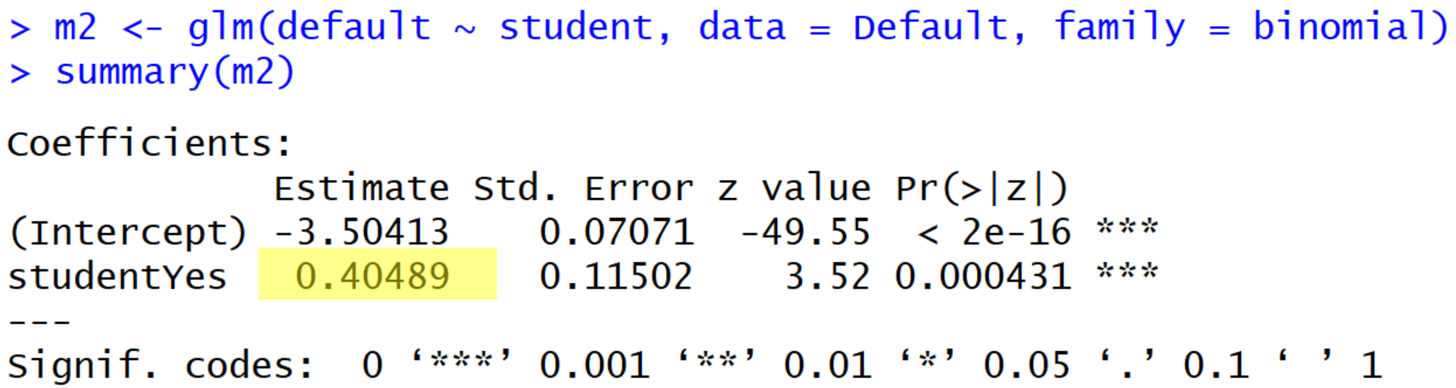


```
> m1 <- glm(default ~ 1, data = Default, family = binomial)
> summary(m1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.36833    0.05574  -60.43   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- $\beta_0 = -3.36833$ indicates:
  - The overall odds of default is $\exp(-3.36833) = 0.034447$
  - The overall probability of default $p = \frac{\exp(-3.36833)}{1+\exp(-3.36833)} = 0.0333$

Logistic Regression with one binary predictor

- Add a binary predictor *student (yes, no)* to the null model
  - StudentYes = 1 or 0
- Model: $\log(\frac{p}{1-p}) = \beta_0 + \beta_1 \mathrm{StudentYes}$

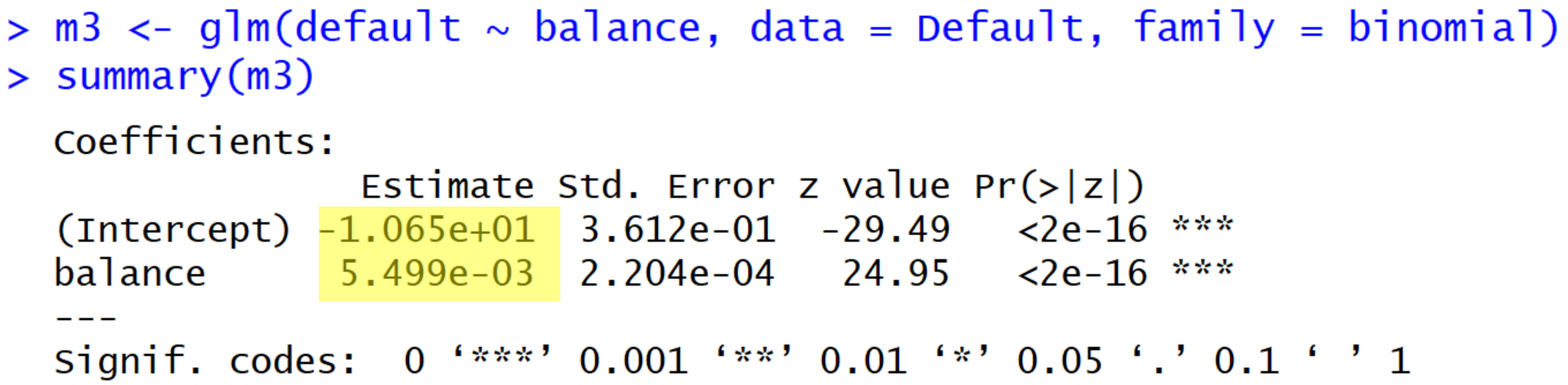


```
> m2 <- glm(default ~ student, data = Default, family = binomial)
> summary(m2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
studentYes   0.40489    0.11502    3.52 0.000431 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- What does the coefficient $\beta_1 = 0.40489$ mean?
  - The log of odds of default for a student is 0.40489 higher than that for a non-student
  - The odds ratio between students and non-students is $\exp(0.40489) = 1.499$
  - The odds (of default) for students is about 50% higher than the odds for non-students.
  - It says nothing general about the probability of default, since evaluating the probability involves not only β_1 but also β_0
- What does the coefficient $\beta_0 = -3.50413$ mean?

Logistic Regression with one continuous variable

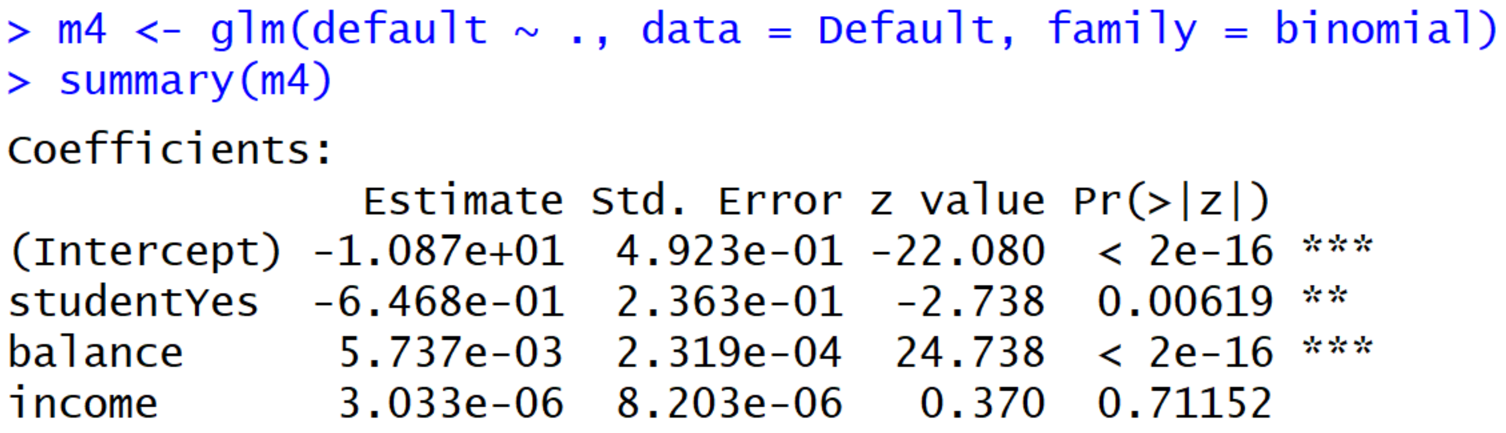
- Add a continuous predictor *balance* to the null model
- Model: $\log(\frac{p}{1-p}) = \beta_0 + \beta_1 \text{Balance}$



```
> m3 <- glm(default ~ balance, data = Default, family = binomial)
> summary(m3)
 Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
 (Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
 balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
 ---
 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

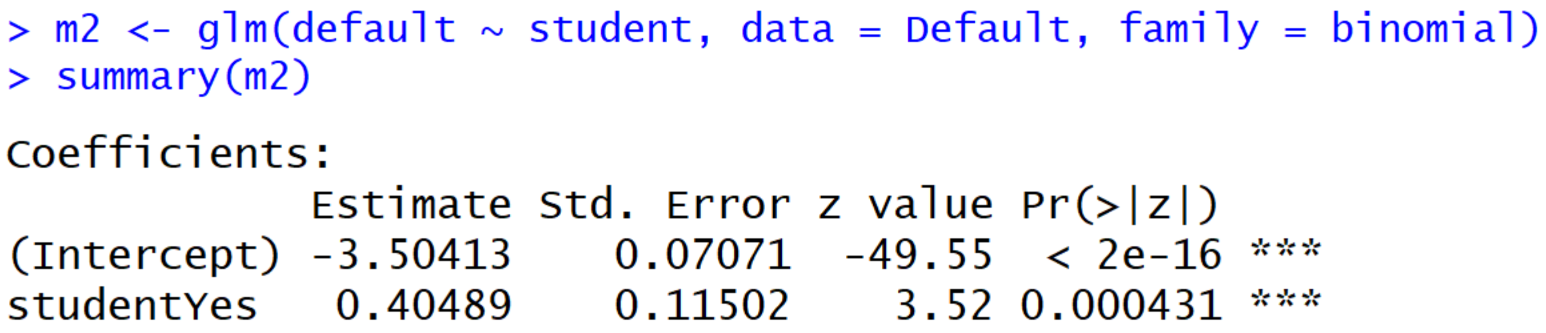
- Let’s fix Balance to 1000, then the conditional logit of default given Balance = 1000 is $-10.65 + 0.005499 \times 1000 = -5.151$
- The odds of default is $\exp(-5.151) = 0.0058$
- The probability of default is $\frac{0.0058}{1+0.0058} = 0.00577$
- A one dollar increase in Balance will (regardless of the current Balance value):
  - Increase the logit of default by 0.005499
  - $\log\left(\frac{p}{1-p}\right)_{after} - \log\left(\frac{p}{1-p}\right)_{before} = \log\left(\frac{\text{odds}_{\text{after}}}{\text{odds}_{\text{before}}}\right) = 0.005499$
  - $\frac{\text{odds}_{\text{after}}}{\text{odds}_{\text{before}}} = \exp(0.005499) = 1.0055$, i.e., increase the odds by 5.5 permille

Multiple Logistic Regression



```
> m4 <- glm(default ~ ., data = Default, family = binomial)
> summary(m4)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 ** 
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152    
```

- What does the negative coefficient on *studentYes* mean?
  - For a fixed value of balance and income, a student is less likely to default than a non-student.



```
> m2 <- glm(default ~ student, data = Default, family = binomial)
> summary(m2)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
studentYes   0.40489    0.11502    3.52 0.000431 ***
```

- With no information about balance and income, a student is more likely to default than a non-student.

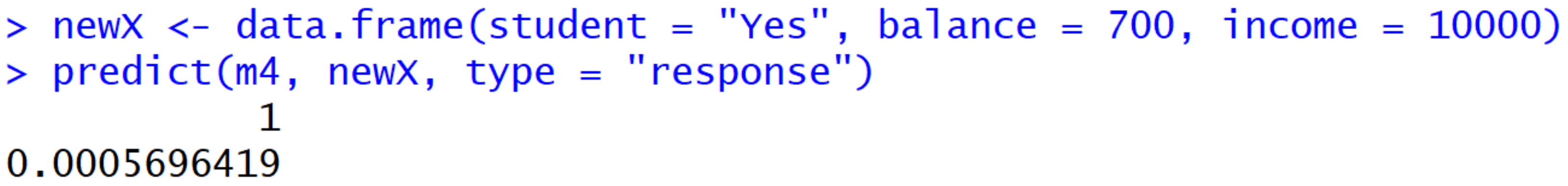
Brain teaser

(reportedly a Microsoft interview question for the Data Analyst position)

- Say there is such a country in which if a boy is born, the parents are happy and stop having more babies, if a girl is born, the parents will keep giving birth to babies until a boy is born.
- Question: in the long run, will there be more males than females or the other way around in this country?

Making predictions

- Once you have a logistic regression model, given a new case X, you can calculate the event probability, or score



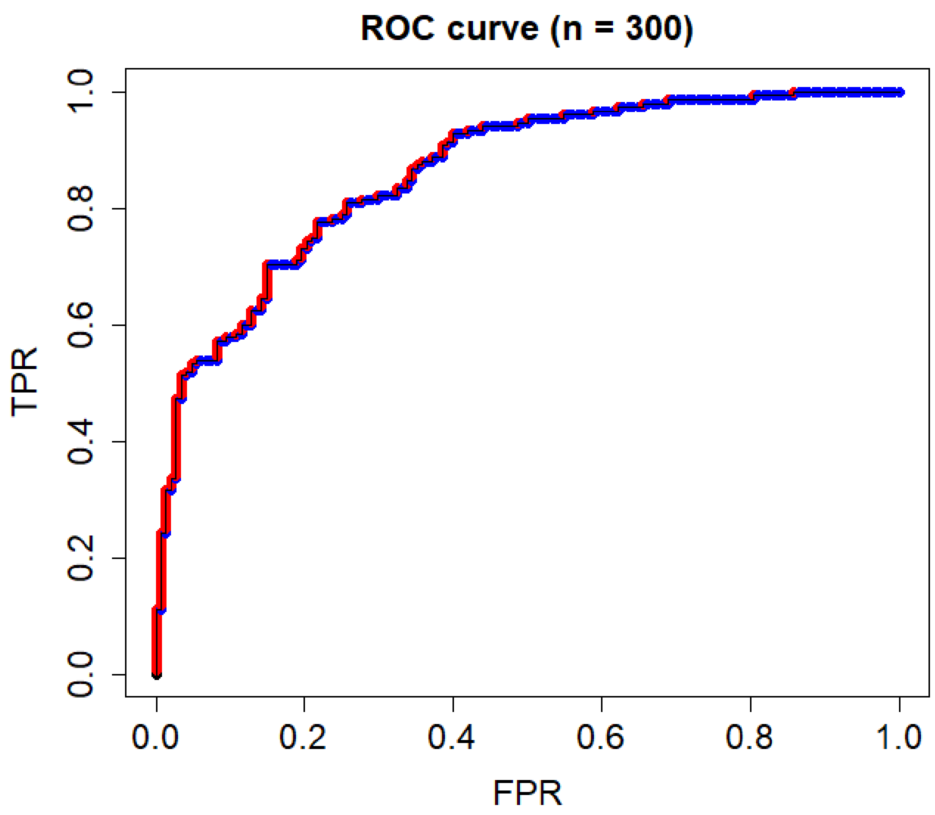
```
> newX <- data.frame(student = "Yes", balance = 700, income = 10000)
> predict(m4, newX, type = "response")
           1
0.0005696419
```

- How would you classify this case? Is this person going to be a Defaulter or a Non-Defaulter?
- You need a score threshold. If the score is higher than the threshold, classify the case as 1 (or positive); otherwise, classify as 0 (or negative).
- Because the model is not perfect, at any threshold the model is bound to make some misclassifications (or false claims), i.e., falsely claiming positive cases as negative, and/or falsely claiming negative cases as positive

ROC curve

- The ROC (Receiver Operating Characteristic) curve is a widely used performance measure for binary classifiers
- It summarizes the effectiveness of the model itself (i.e., at all possible classification thresholds), rather than the effectiveness of a particular classification rule (i.e., at a particular classification threshold).
- Choice of the threshold is a business decision, not a model property
  - How much cost is associated with a false positive and false negative prediction, respectively?
  - How much benefit is associated with a true positive and true negative prediction, respectively?
  - The answer varies greatly by domain, e.g., medical, marketing, …
- ROC can help determine the appropriate threshold, since it plots out the classification performance at all possible thresholds

TPR and FPR



Classification Rule:

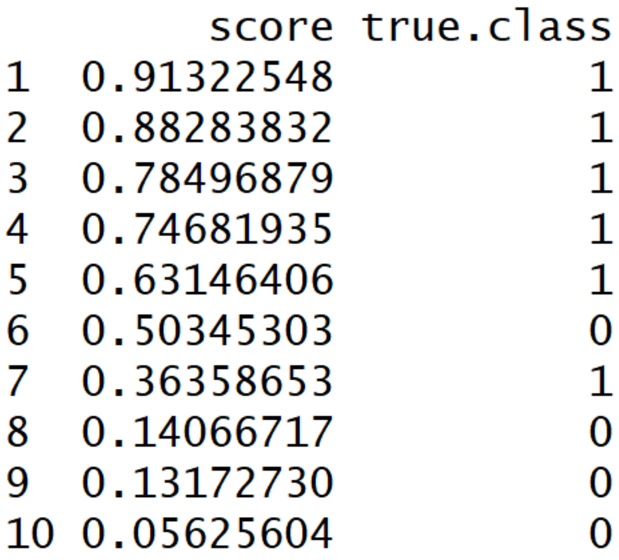
- If the case score is greater than the threshold, classify the case as positive;
- If the case score is no greater than the threshold, classify the case as negative.

At any threshold, a FPR and a TPR can be calculated.

False Positive Rate (FPR): The number of cases that the model classifies as positive which are actually negative divided by the total number of negative cases in the data set.
True Positive Rate (TPR): The number of cases that the model classifies as positive which are indeed positive divided by the total number of positive cases in the data set.

Construct the ROC curve

- Given a test set of n cases, say there are P positives and N negatives, with $P + N = n$
- Calculate the score using the model (using the predict() function)
- Sort the cases by score in descending order (high to low)



```
         score true.class
1   0.91322548          1
2   0.88283832          1
3   0.78496879          1
4   0.74681935          1
5   0.63146406          1
6   0.50345303          0
7   0.36358653          1
8   0.14066717          0
9   0.13172730          0
10  0.05625604          0
```

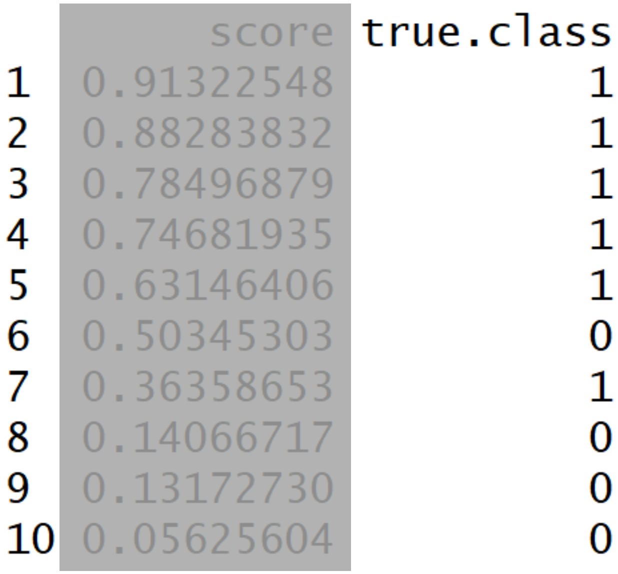
Place pencil tip at (0,0) in an empty TPR-FPR graph, pencil tip not to leave the paper until all done.

Start from the top of the sorted list (in fact, after sorting, the score values do not matter any more), read the true.class value row by row:

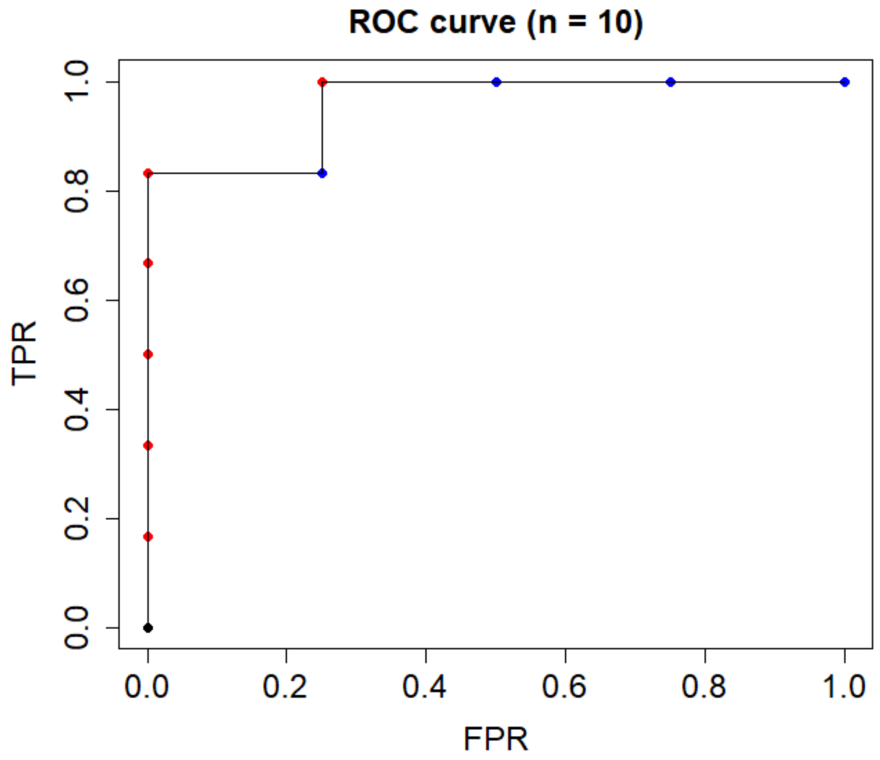
- If you see a 1, move pencil tip upward by a length of 1/P
- If you see a 0, move pencil tip rightward by a length of 1/N

Until the end of the list, at which point the pencil tip should be at coordinate (1,1). The ROC curve is complete.

Construct the ROC curve



```
       score true.class
1  0.91322548          1
2  0.88283832          1
3  0.78496879          1
4  0.74681935          1
5  0.63146406          1
6  0.50345303          0
7  0.36358653          1
8  0.14066717          0
9  0.13172730          0
10 0.05625604          0
```

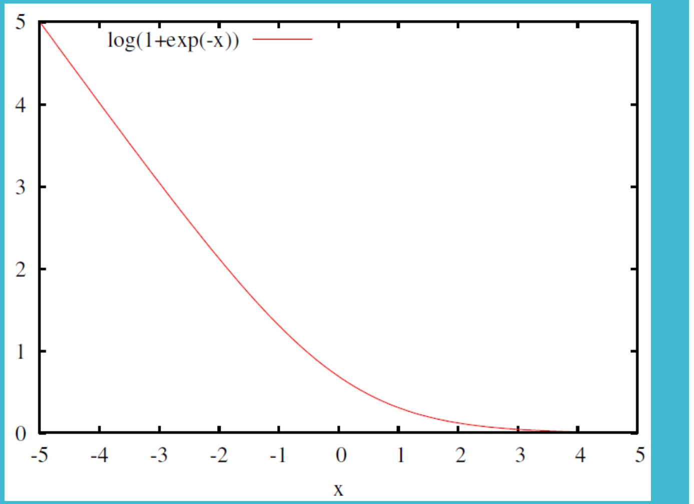


https://www.linkedin.com/pulse/roc-curve-simple-terms-yanchao-liu/

Area Under the Curve (AUC)

- AUC typically ranges between 0.5 and 1
- The AUC value is the percentage of randomly drawn pairs (one from positive population and the other from the negative population) for which the model scores correctly discriminate the positive and negative case in the pair
  - "Correctly discriminate" means that the positive case receives a higher score than the negative case
- AUC represents the overall model performance, agnostic of how the actual classification is done

Solving for the coefficients (optional)

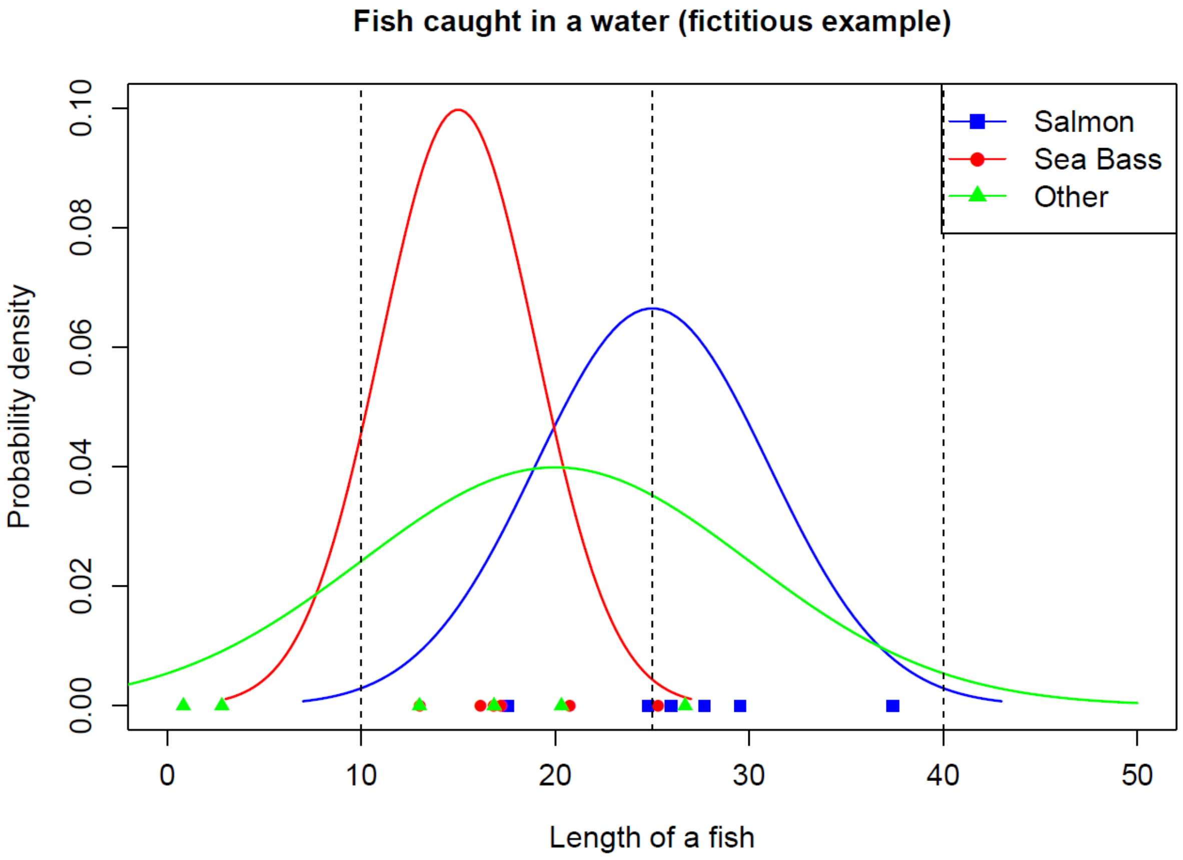


- Given a set of training samples $\{(X_i, y_i)\}_{i=1}^{n}$
  - Encode event as $y = 1$ and non-event as $y = -1$
- For an event, we want $(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})$ to be positive and as large as possible
- For a non-event, we want $(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})$ to be negative and its absolute value as large as possible
- Overall, we want $a_i \coloneqq y_i(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})$ to be as large as possible for all each training case i.
- We solve for the coefficients $\beta_j, j = 0, \ldots, p$ by penalizing a smooth decreasing function of a_i, the log loss function: $\log(1 + \exp(-a_i))$
- Minimize $\sum_{i=1}^{n} \log(1 + \exp(-y_i\left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\right)))$

Discriminant Analysis

- Applicable when Y is categorical and X is continuous
- For example, use the length and lightness to predict the type of a fish
- Mathematical machinery:
  - Bayes Theorem
  - Gaussian distribution
- Method: construct a discriminant function for each class, classify a new input X as the class for which the discriminant function is the largest.
  - Linear Discriminant Analysis (LDA): the discriminant function is a linear function of X
  - Quadratic Discriminant Analysis (QDA): the discriminant function is a quadratic function of X

Definition of π_k and $f_k(x)$



X = Length of fish
Y = Type of fish

We wish to classify fish caught in a water into $K = 3$ classes

Let π_k represent the overall or prior probability that a randomly chosen observation comes from the class k

- $\pi_{\text{Salmon}} = 0.3, \pi_{\text{SeaBass}} = 0.6, \pi_{\text{Other}} = 0.1$

Let $f_k(x) \coloneqq \Pr(X = x \mid Y = k)$ denote the density function of X for an observation that comes from class k.

Bayes Theorem

Posterior probability of the event $Y = k$

- Bayes' theorem states that $p_k(x) := \Pr(Y = k | X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$
- Bayes' decision rule: classify x to be in class k for which $p_k(x)$ is largest
- Example:
  - $\pi_{\text{Salmon}} = 0.3, \pi_{\text{SeaBass}} = 0.6, \pi_{\text{Other}} = 0.1$
  - $X \sim Normal(20, 5)$ for $Y = \text{Salmon}$
  - $X \sim Normal(12, 2.4)$ for $Y = \text{SeaBass}$
  - $X \sim Normal(15, 10)$ for Other
  - For $x = 10$, what are $p_k(x)$ for $k =$ Salmon, SeaBass and Other, respectively?
  - How would you classify a fish with $x = 10$?

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

Simplify calculations

- The k for which $p_k(x)$ is largest is the same as the k for which $\pi_k f_k(x)$ is largest, because $p_k(x)$ for all $k = 1, \dots K$ share the same denominator.
- If we assume that $f_k(x)$ are normal densities with parameters μ_k and σ
- Then, that $\pi_k f_k(x)$ is largest is equivalent to that

$$x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

is largest, where $\hat{\mu}_k$, $\hat{\sigma}$ and $\hat{\pi}_k$ are estimated from training data.

- This *discriminant function*, denoted by $\delta_k(x)$, is a linear function of the input x, therefore the method is called **Linear Discriminant Analysis** (LDA).

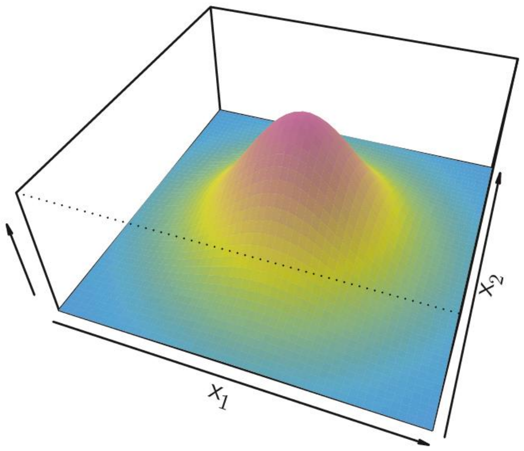
Exercise (LDA)

	length	species
1	18.7	Salmon
2	15.4	Salmon
3	19.8	Salmon
4	20.8	Salmon
5	25.2	Salmon
6	19.2	Salmon
7	18.3	SeaBass
8	14.6	SeaBass
9	10.2	SeaBass
10	13.9	SeaBass
11	10.1	SeaBass
12	8.9	SeaBass
13	14.1	SeaBass
14	10.7	SeaBass
15	12.7	SeaBass
16	8.8	SeaBass
17	6.5	SeaBass
18	15	Other
19	9.9	Other
20	15.6	Other

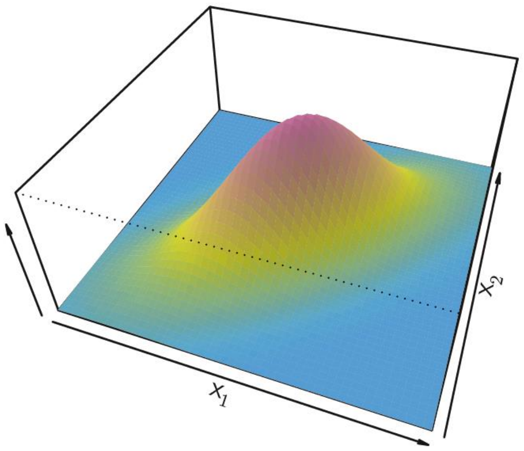
- Your friend caught 20 fish from a water, labeled their species and measured their lengths. Data is shown on the left.
- You are new to fishing and happily catches your first fish in the same water, and you are eager to know its species.
- The fish is 15 inches long.
- Questions: Use LDA to predict its species
  - State the assumptions needed for LDA
  - Write out the discriminant function and state the decision rule
  - Estimate the needed parameters
  - Calculate the discriminant function values and draw a conclusion, what species is the fish?

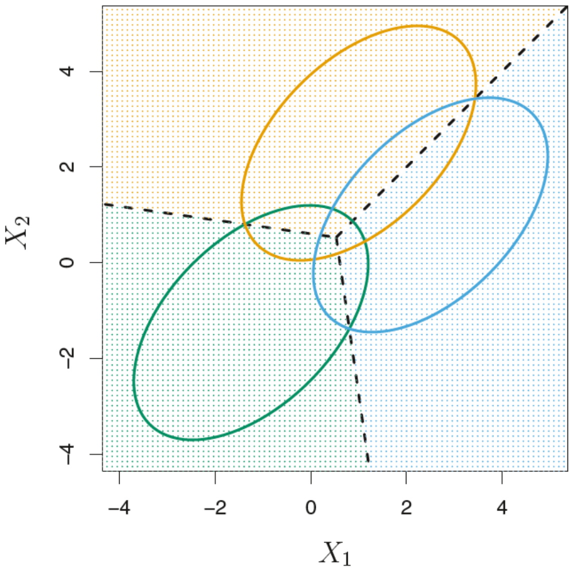
LDA for $p > 1$

X1 and X2 uncorrelated



X1 and X2 correlated, $\rho = 0.7$





LDA Assumptions for $x = (x_1, \dots, x_p)$:

- Each $f_k(x)$ is a multi-variate Gaussian density with μ_k and a common covariance matrix Σ

The rings are at different locations but of the same shape

LDA for $p > 1$

- The discriminant function becomes

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

- Example $p = 2$, $x = (x_1, x_2)$ = (Length, Lightness)
  - Suppose from training data you have estimated
    - $\mu_{\text{Salmon}} = \begin{bmatrix} 20 \\ 4 \end{bmatrix}$, $\mu_{\text{SeaBass}} = \begin{bmatrix} 12 \\ 6 \end{bmatrix}$, $\mu_{\text{Other}} = \begin{bmatrix} 15 \\ 2 \end{bmatrix}$
    - $\Sigma = \begin{bmatrix} 9 & -2 \\ -2 & 1 \end{bmatrix}$, $\pi_k = \begin{bmatrix} 0.6 \\ 0.3 \\ 0.1 \end{bmatrix}$
  - For a new fish of length 15 and lightness 3.5, calculate $\delta_k(x)$ for k = Salmon, SeaBass and Other, and make a prediction.

Bayes Theorem

- Bayes Theorem: $\mathrm{P}(X \ and \ Y) = \mathrm{P}(X \mid Y)\mathrm{P}(Y) = \mathrm{P}(Y \mid X)\mathrm{P}(X)$
  - $P(Y = k \mid X = x) = P(Y = k) * P(X = x \mid Y = k) \ / \ P(X = x)$
  - Posterior = Prior * Likelihood / Evidence
  - $P(X = x) = \sum_{k=1}^{K} P(X = x \mid Y = k) * P(Y = k)$

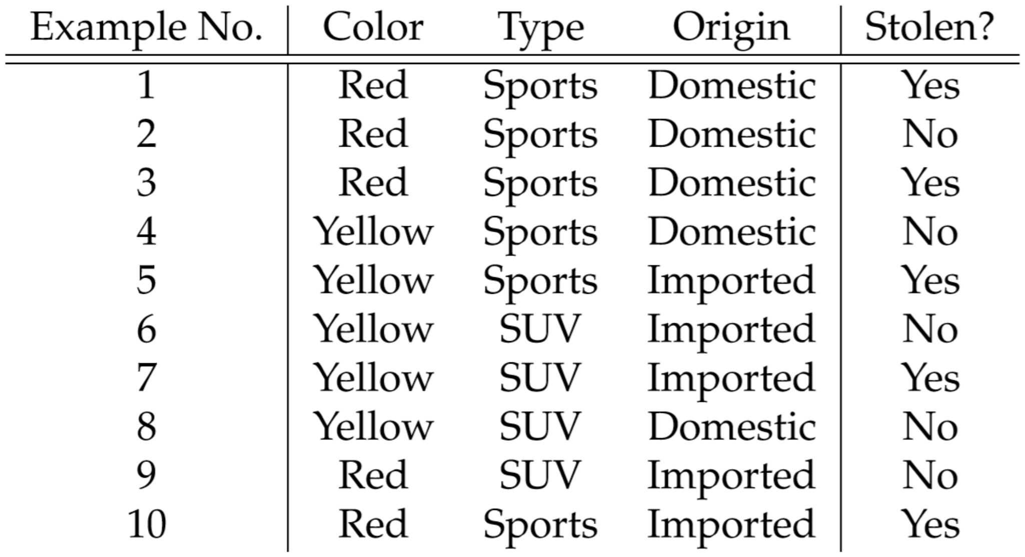
- 99% of those who are affected by a virus have a positive result in a virus test. Someone got a positive test result, should she be scared?

- In addition, only 1% of those who are tested actually get a positive result. Should she now be scared?

Naïve Bayes Classifier

- When X is multi-dimensional (i.e., $p > 1$), how do we calculate $P(X = x \mid Y = k)$?
  - **LDA** assumes that X given $Y = k$ is a multi-variate Gaussian random vector with mean vector μ_k and a common covariance matrix Σ
  - **Naïve Bayes** assumes X_j's are independent random variables given Y
    - Best applicable when X_j's are categorical variables
- Naïve Bayes: assume conditional independence among x_j
  - $P\big((X_1, \ldots X_p) = (x_1, \ldots, x_p) \mid Y = k\big) = P(X_1 = x_1 | \, Y = k) * \cdots * P\big(X_p = x_p \mid Y = k\big)$
- So overall, $P(Y = k \mid X = x) = \frac{P(Y=k)*P(X_1=x_1 | \, Y=k)*\cdots*P\left(X_j=x_j \middle| \, Y=k\right)}{P(X=x)}$
- Predict x to be in class k for which $P(Y = k \mid X = x)$ is largest.
- Only need to calculate and compare the numerators, since the denominator is the same for all k

Naïve Bayes Example



Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

http://www.inf.u-szeged.hu/~ormandi/ai2/06-naiveBayes-example.pdf

We want to classify a Red Domestic SUV. Show the calculation steps and make a conclusion.