

THE UNIVERSITY OF TEXAS AT AUSTIN



Classification:  
**Logistic Regression and Naive Bayes**  
Book Chapter 4.

**Jared S. Murray**  
The University of Texas McCombs School of Business

1. Classification
2.  $k$ -Nearest Neighbors (kNN)
3. Classification Trees
4. Logistic Regression, One Predictor
5. Inference: Estimating the Parameters
6. Multiple Logistic Regression
7. AIC and BIC in Logistic Regression
8. Making Decisions and Loss
9. Out of Sample
10. Lift and ROC
11. Bayes Theorem and Classification
  - 11.1. Quick Basic Probability Review
  - 11.2. Conditional Probability and Classification
  - 11.3. Bayes Theorem
  - 11.4. Naive Bayes

## 1. Classification

When the  $y$  we are trying to predict is *categorical* (or *qualitative*), we say we have a *classification* problem.

For a numeric (or *quantitative*)  $y$  we predict it's value.

For a categorical  $y$  we try to guess which of a listed number of possible outcomes will happen.

The basic case is a binary  $y$ : something will either happen or not.

# Classification

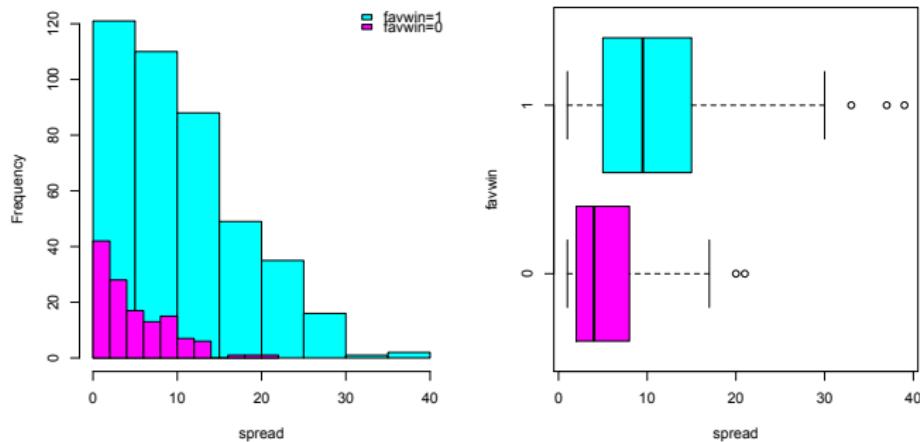
We've mentioned the classification versions of knn and trees (in passing).

The main difference between regression and classification is how we measure error/accuracy, and how we tune our models/methods.

In this section of notes we will discuss this and introduce two new methods: logistic regression and naive Bayes.

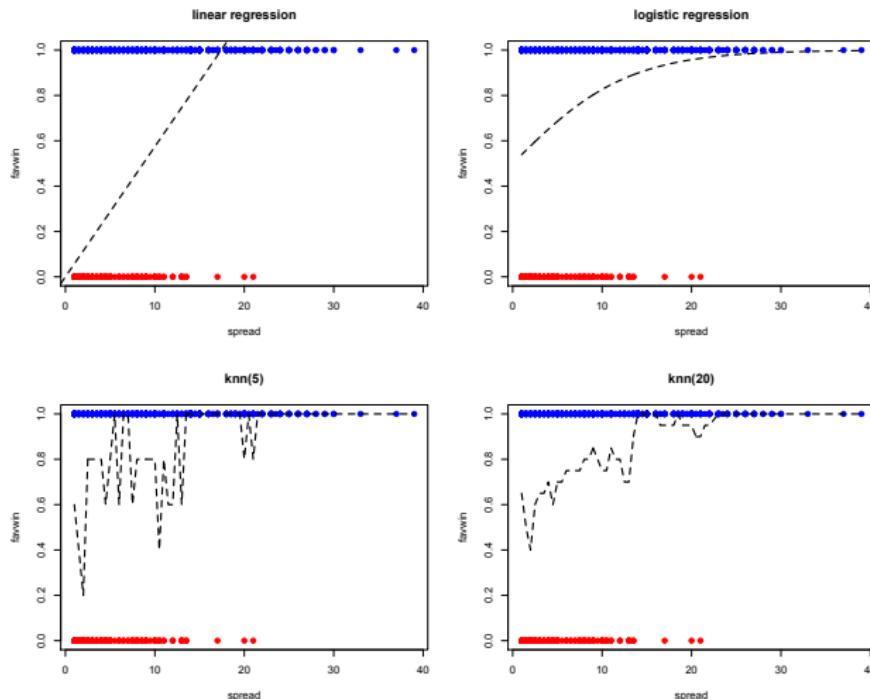
# Classification

Below is an example where we are trying to predict whether or not the favorite team will win as a function of the Vegas' betting point-spread.



# Classification

In this context,  $f(X)$  will output the probability of  $Y$  taking a particular category (win/lose, here). Below, the black line represent different methods to estimate  $f(X)$  in this example.



## Classification

All the discussion about complexity vs. predictive accuracy (bias-variance trade-off) applies to this setting... what differs is our measure of accuracy as we are no longer dealing with a numerical outcome variable. A common approach is to measure the *error rate*, i.e., the number of times we a estimated  $\widehat{f(\cdot)}$  “gets it wrong” ... in the training set (with  $n$  observations) define the error rate as:

$$\frac{1}{n} \sum_{i=1}^n I(Y_i \neq \widehat{Y}_i)$$

where  $\widehat{Y}_i$  is the class (category) label assigned by  $\widehat{f(X_i)}$  and  $I(\cdot)$  is an indicator variable that equals 1 whenever  $Y_i \neq \widehat{Y}_i$  and 0 otherwise.

## Classification

As mentioned above, in classification,  $\widehat{f(X_i)}$  outputs a probability. In order to use this information and create a *classification rule* we need to decide on a **cut-off** (or cut-offs) to determine the label assignments.

In general, if we are looking at only two categories (like the NBA example above) the standard cut-off is 0.5. This means that if  $\widehat{f(X_f)} > 0.5$  we define  $\hat{Y}_i = 1$  ( $\hat{Y}_i = 0$  otherwise).

Later on, we will think more carefully about defining cut-offs...

## 2. *k*-Nearest Neighbors (kNN)

The *k*-nearest neighbors algorithm will try to *predict* (numerical variables) or *classify* (categorical variables) based on *similar (close) records* on the *training dataset*.

Remember, the problem is to guess a future value  $Y_f$  given new values of the covariates  $X_f = (x_{1f}, x_{2f}, x_{3f}, \dots, x_{pf})$ .

## *k*-Nearest Neighbors (kNN)

**kNN:** How do the  $Y$ 's look like close to the region around  $X_f$ ?

We need to find the  $k$  records in the training dataset that are close to  $X_f$ . How? “Nearness” to the  $i^{th}$  neighbor can be defined by (euclidian distance):

$$d_i = \sqrt{\sum_{j=1}^p (x_{jf} - x_{ji})^2}$$

### **Prediction:**

- ▶ Numerical  $Y_f$ : take the average of the  $Y$ 's in the  $k$ -nearest neighbors
- ▶ Categorical  $Y_f$ : take the most common category in the  $k$ -nearest neighbors

# *k*-Nearest Neighbors (kNN) – Example

## Forensic Glass Analysis



Classifying shards of glass

Refractive index, plus oxide %

Na, Mg, Al, Si, K, Ca, Ba, Fe.

6 possible glass types

WinF: float glass window

WinNF: non-float window

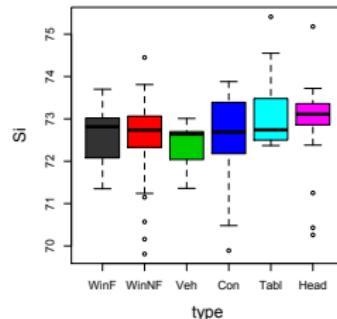
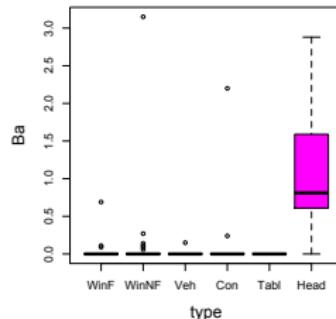
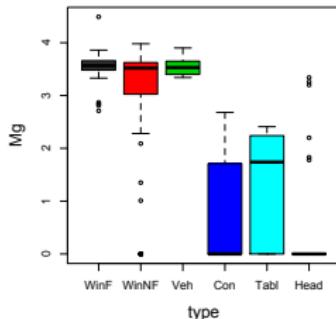
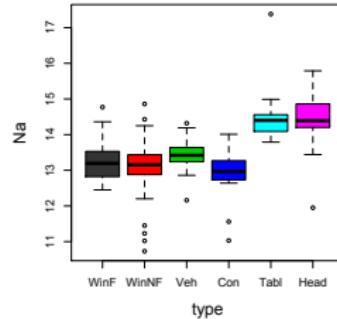
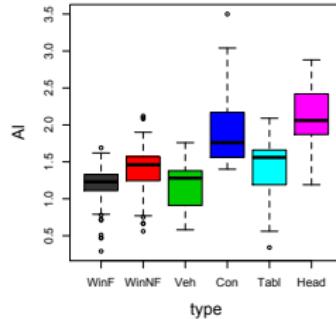
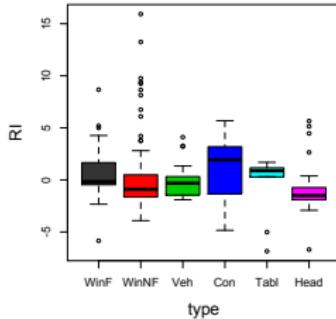
Veh: vehicle window

Con: container (bottles)

Tabl: tableware

Head: vehicle headlamp

# $k$ -Nearest Neighbors ( $k$ NN) – Example

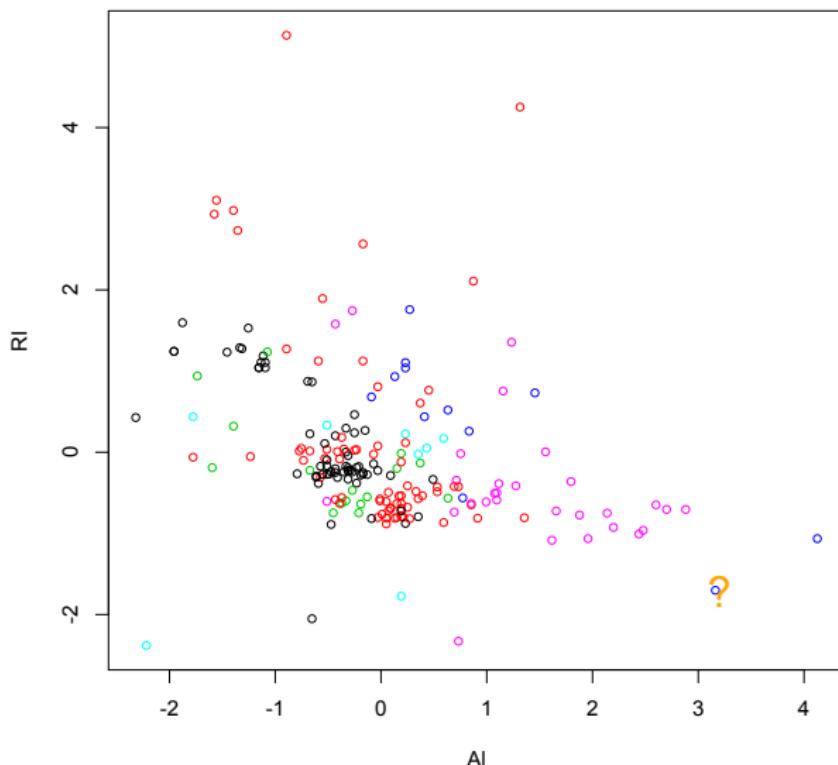


Some variables are clear discriminators (Ba) while others are more subtle (RI)

## $k$ -Nearest Neighbors ( $k$ NN) – Example

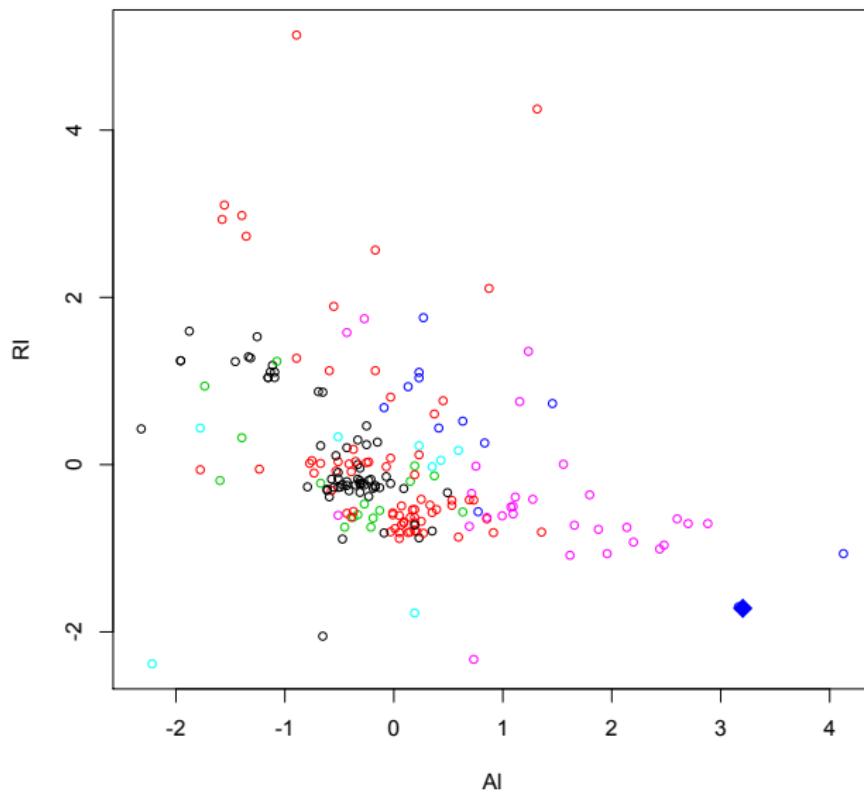
Using only RI and AI let's try to predict the point marked by “?”

1-nearest neighbor



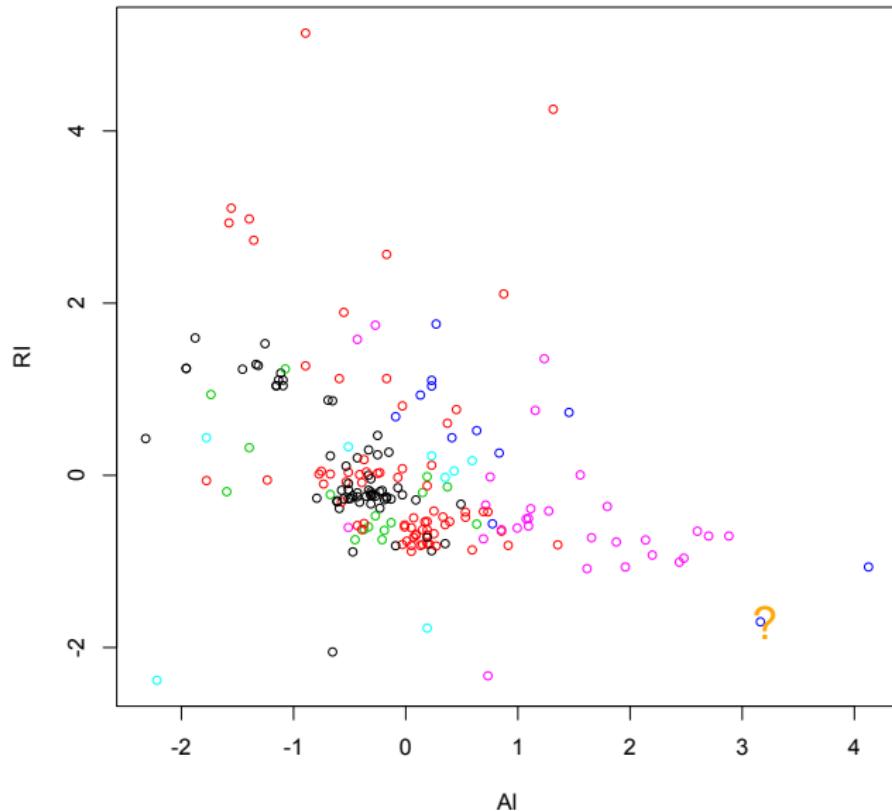
# $k$ -Nearest Neighbors ( $k$ NN) – Example

1-nearest neighbor



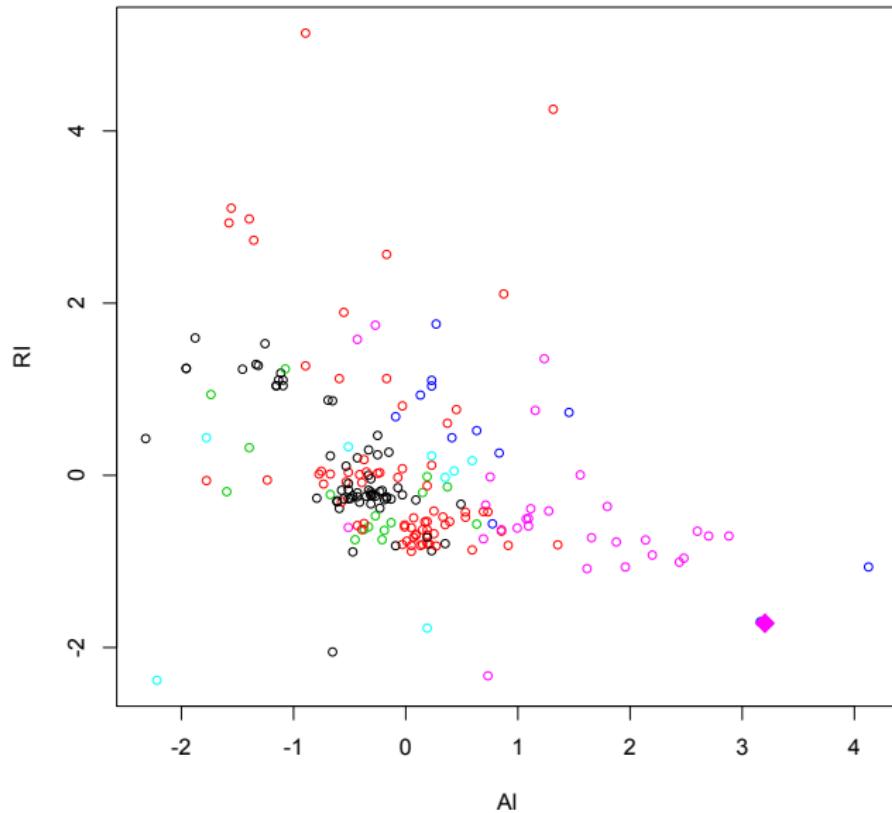
# $k$ -Nearest Neighbors ( $k$ NN) – Example

5-nearest neighbor



# $k$ -Nearest Neighbors ( $k$ NN) – Example

5-nearest neighbor



## *k*-Nearest Neighbors (kNN)

### **More comments:**

- ▶ The distance metric used above is only valid for numerical values of  $X$ . When  $X$ 's are categorical we need to think about a different distance metric or perform some manipulation of the information.
- ▶ The scale of  $X$  also will have an impact. In general it is a good idea put the  $X$ 's in the same scale before running kNN (see example in R)

### 3. Classification Trees

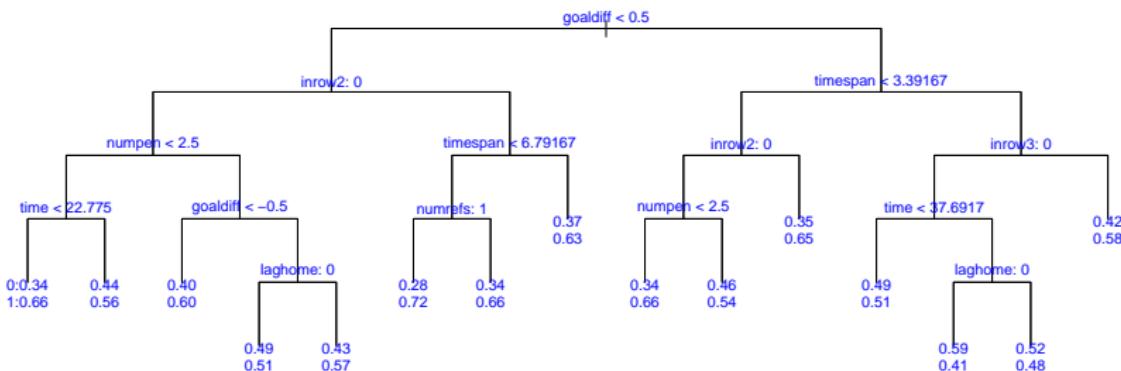
Let's do a tree for a classification problem.

We'll use the hockey penalty data.

The response is whether or not the next penalty is on the other team and  $x$  is a bunch of stuff about the game situation (the score, etc ...).

In addition, this time some of our predictors (features,  $x$ 's) are categorical.

Here is the tree:



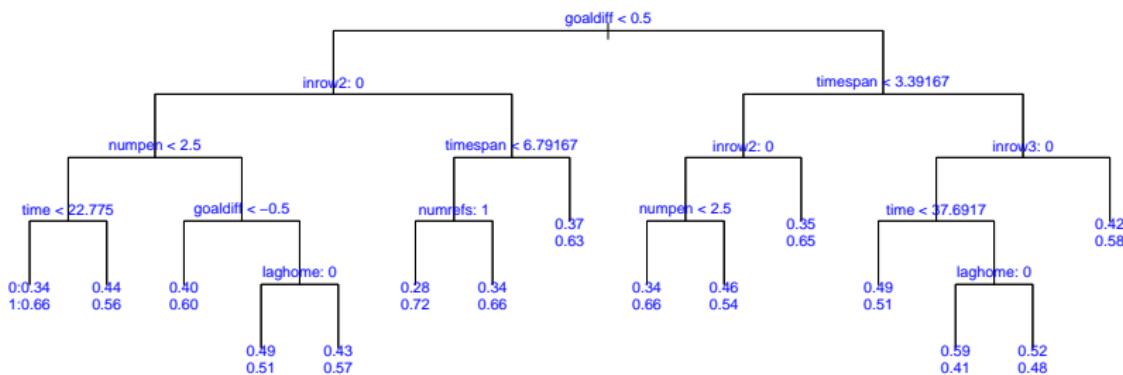
- ▶ Each bottom node gives the fraction of training data in the two outcome categories. Think of it as  $\hat{p}$  for the kind of  $x$  associated with that bottom node.
- ▶ The form of the decision rule can't be  $x < c$  for categorical variables. We pick a subset of the levels to go left.  $inrow2:0$  means all the observations with  $inrow2$  in the category labeled 0 go left.

*if:*

- ▶ if you are not winning
- ▶ you had the last two penalties
- ▶ it has not been long since the last call
- ▶ and there is only 1 referee

*then:*

there is a 72% chance the next call will be on the other team.



Whilst there is another game situation where the chance the next call is on the other team is only 41%.

## 4. Logistic Regression, One Predictor

To start off as simply as possible, we will first consider the case where we have a binary  $y$  and one numeric  $x$ .

Lets' look at the Default data (from Chapter 4):

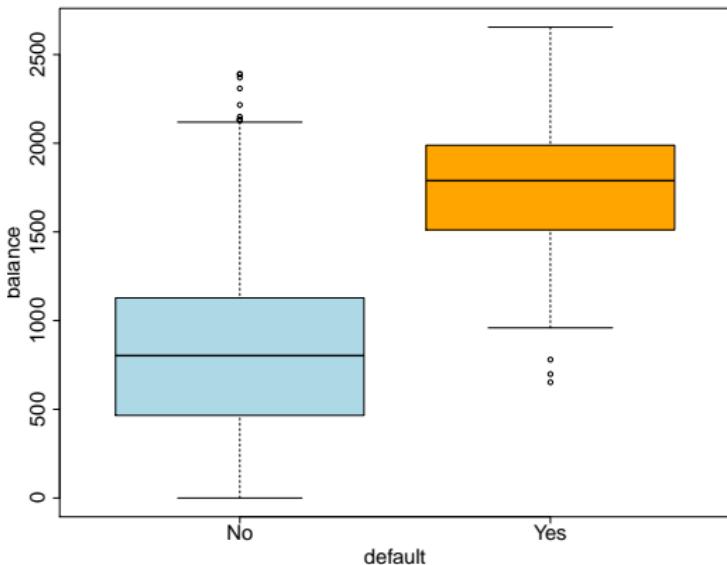
- ▶  $y$ :  
whether or not a customer defaults on their credit card  
(No or Yes).
- ▶  $x$ :  
The average balance that the customer has remaining on their credit card after making their monthly payment.
- ▶ 10,000 observations, 333 defaults (.0333 default rate).

Let's look at the data.

Divide the data into two groups, one group has  $y=\text{No}$  and other other group has  $y=\text{Yes}$ .

Use boxplots to display the  $x=\text{balance}$  values in each subgroup.

The balance values are bigger for the default  $y=\text{Yes}$  observations!



In our example, we would want

$$Pr(Y = Yes \mid X = x).$$

Given the probability we can classify using a rule like

guess Yes if:  $Pr(Y = Yes \mid x) > .5.$

## Notation:

For a binary  $y$ , it is very common to use a dummy variable to code up the two possible outcomes.

So, in our example, we might say a default means  $Y = 1$  and a non-default means  $Y = 0$ .

In the context of an example we might use the label and  $Y = 1$  interchangeably.

In our default example,  $P(Y = 1 | X = x)$  and  $P(Y = \text{yes} | X = x)$  would mean the same thing.

Normally, we might use names like  $D$  and  $B$  for our two variables, but since we want to think about the ideas in general, let's stick with  $Y$  and  $X$ .

Logistic regression uses the power of linear modeling and estimates  $Pr(Y = y | x)$  by using a two step process.

► Step 1:

apply a linear function to  $x$ :  $x \rightarrow \eta = \beta_0 + \beta_1 x$ .

► Step 2:

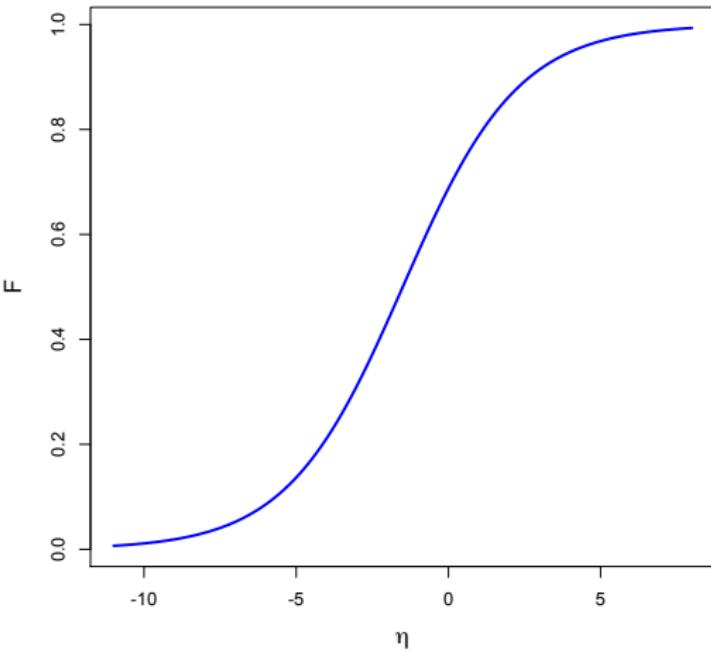
apply the *logistic function*  $F$ ,  
to  $\eta$  to get a number between 0 and 1.  
 $P(Y = 1 | x) = F(\eta)$ .

## The logistic function:

$$F(\eta) = \frac{e^\eta}{1 + e^\eta}.$$

The key idea is that  $F(\eta)$  is always between 0 and 1 so we can use it as a probability.

Note that  $F$  is increasing, so if  $\eta$  goes up  $P(Y = 1 | x)$  goes up.



$$F(-3) = .05, F(-2) = .12, F(-1) = .27, F(0) = .5$$

$$F(0) = .5$$

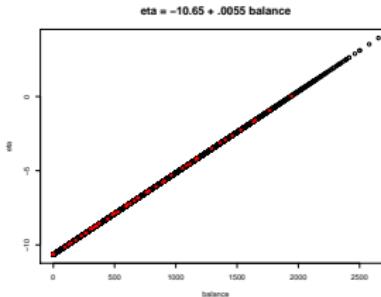
$$F(1) = .73, F(2) = .88, F(3) = .95$$

## Logistic fit to the $y=\text{default}$ , $x=\text{balance}$ data.

First, logistic looks for a linear function of  $x$  it can feed into the logistic function.

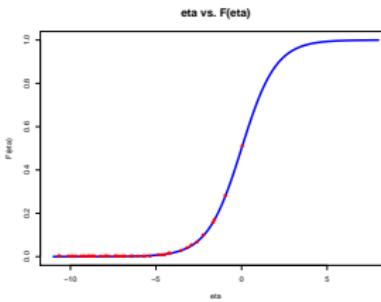
Here we have

$$\eta = -10.65 + .0055 x.$$



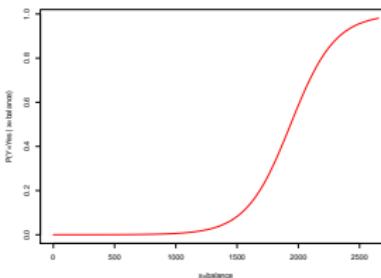
Next we feed the  $\eta$  values into the logistic function.

100 randomly sampled observations are plotted with red dots.



We can combine the two steps together and plot  $x=\text{balance}$  vs.

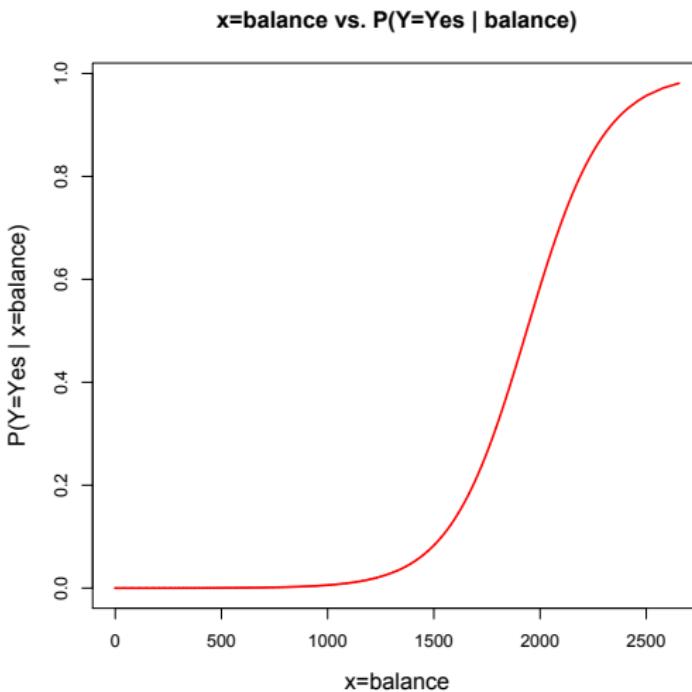
$$P(Y = \text{Yes} | x) = F(-10.65 + .0055 x).$$



## Logistic Regression:

Combining the two steps, our logistic regression model is:

$$P(Y = 1 \mid X = x) = F(\beta_0 + \beta_1 x).$$



## 5. Inference: Estimating the Parameters

Logistic regression gives us a formal parametric statistical model (like linear regression with normal errors).

Our model is:

$$Y_i \sim \text{Bernoulli}(p_i), \quad p_i = F(\beta_0 + \beta_1 x_i).$$

Our model has two parameters  $\beta_0$  and  $\beta_1$  which can estimate given data.

We usually assume that the given the parameters and  $x_i$ , the  $Y_i$  are independent.

To estimate the parameters, we usually use *maximum likelihood*.

That is, we choose the parameter values that make the data we have seen most likely.

Let  $p_y$  be a simplified notation for  $P(Y = y | x)$ .

In the logistic model,  $p_y$  depends on  $(\beta_0, \beta_1)$

$$p_y = p_y(\beta_0, \beta_1) = \begin{cases} P(Y = 1 | x) = F(\beta_0 + \beta_1 x) & Y = 1 \\ P(Y = 0 | x) = 1 - F(\beta_0 + \beta_1 x) & Y = 0 \end{cases}$$

For our logistic model, the probability of the  $Y_i = y_i$  given  $x_i$ ,  $i = 1, 2, \dots, n$  as a function of  $\beta_0$  and  $\beta_1$  is

$$L(\beta_0, \beta_1) = \prod_{i=1}^n p_{y_i}(\beta_0, \beta_1).$$

So, we estimate  $(\beta_0, \beta_1)$  by choosing the values that optimize the likelihood function  $L(\beta_0, \beta_1) !!$

This optimization has to be done numerically using an iterative technique (Newton's Method!!).

The problem is convex and the optimization usually converges pretty fast.

Some fairly complex statistical theory gives us standard errors for our estimates from which we can get confidence intervals and hypothesis test for  $\beta_0$  and  $\beta_1$ .

Here is the logistic regression output for our  $y=\text{default}$ ,  $x=\text{balance}$  example.

The MLE of  $\beta_0$  is  
 $\hat{\beta}_0 = -10.65$ .

The MLE of  $\beta_1$  is  
 $\hat{\beta}_1 = .0055$ .

Given  $x=\text{balance} = 2000$ ,  
 $\eta = -10.65 + .0055 * 2000 = 0.35$

$\hat{\beta}_1 > 0$  suggests larger balances are associated with higher risk of default.

$P(\text{default}) =$   
 $P(Y = 1 | x = 2000) =$   
 $\exp(.35)/(1+\exp(.35))$   
 $= 0.59$ .

```
Call:  
glm(formula = default ~ balance, family = binomial, data = Default)  
  
Deviance Residuals:  
    Min      1Q  Median      3Q     Max  
-2.2697 -0.1465 -0.0589 -0.0221  3.7589  
  
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  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 2920.6 on 9999 degrees of freedom  
Residual deviance: 1596.5 on 9998 degrees of freedom  
AIC: 1600.5  
  
Number of Fisher Scoring iterations: 8
```

Confidence Interval for  $\beta_1$ :

$$\hat{\beta}_1 \pm 2\text{se}(\hat{\beta}_1).$$

$$.0055 \pm 2(.00022).$$

Test  $H_0 : \beta_1 = \beta_1^0$

$$z = \frac{\hat{\beta}_1 - \beta_1^0}{\text{se}(\hat{\beta}_1)}.$$

If  $H_0$  is true,  $z$  should look like a standard normal draw.

$$\frac{.0055 - 0}{.00022} = 25,$$

big for a standard normal

$\Rightarrow$

reject the null that  $\beta_1 = 0$ .

Similar for  $\beta_0$ .

Call:

```
glm(formula = default ~ balance, family = binomial, data = Default)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.2697	-0.1465	-0.0589	-0.0221	3.7589

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2920.6 on 9999 degrees of freedom

Residual deviance: 1596.5 on 9998 degrees of freedom

AIC: 1600.5

Number of Fisher Scoring iterations: 8

## Fisher Scoring iterations:

It took 8 iterations of the optimization for convergence.

## Deviance:

The deviance is  $-2\log(L(\hat{\beta}_0, \hat{\beta}_1))$ .

Twice (-2) times the log of the maximized likelihood.

For numerical and theoretical reasons it turns out to be easier to work with the log of the likelihood than the likelihood itself.

Taking the log turns all the products into sums.

A big likelihood is good, so a small deviance is good.

## Null and Residual Deviance:

The Residual deviance is just the one you get by plugging the MLE's of  $\beta_0$  and  $\beta_1$  into the likelihood.

The Null deviance is what you get by setting  $\beta_1 = 0$  and then optimizing the likelihood over  $\beta_0$  alone.

You can see that the deviance is a lot smaller when we don't restrict  $\beta_1$  to be 0!!

Deviance as a sum of losses:

If we let

$$\hat{p}_y = p_y(\hat{\beta}_0, \hat{\beta}_1),$$

then the deviance is

$$\sum_{i=1}^n -2 \log(\hat{p}_{y_i}).$$

The sum over observations of -2 times the log of the estimated probability of the  $y$  you got.

Minimizing deviance in logistic regression is equivalent to minimizing *log-loss*...

$p_y$  is the probability we assign to  $Y$  turning out to be  $y$ .

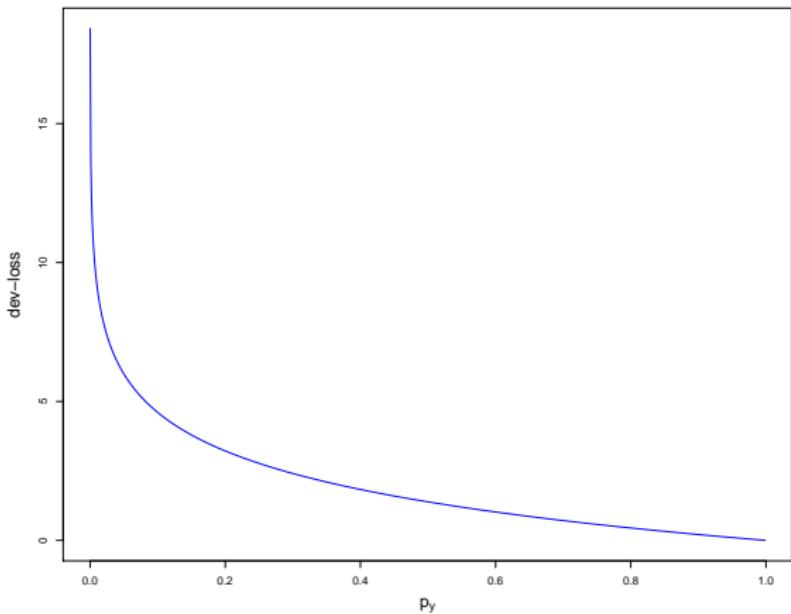
We can think of  $-2 \log(p_y)$  as our *loss*.

This is

$p_y$  versus  $-2 \log(p_y)$ .

When  $y$  happens, the bigger we said  $p_y$  is the better off we are (the lower our loss).

If  $y$  happens and we said  $p_y$  is small, we really get a big loss - that's fair!!



In Section 9 we will use deviance as an *out of sample* loss function, just as we have used sum of squared errors for a numeric  $Y$ .

## 6. Multiple Logistic Regression

We can extend our logistic model to several numeric  $x$  by letting  $\eta$  be a linear combination of the  $x$ 's instead of just a linear function of one  $x$ :

- ▶ Step 1:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p.$$

- ▶ Step 2:

$$P(Y = 1 \mid x = (x_1, x_2, \dots, x_p)) = F(\eta).$$

Or, in one step, our model is:

$$Y_i \sim \text{Bernoulli}(p_i), \quad p_i = F(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}).$$

Our first step keeps some of the structure we are used to in linear regression.

We combine the  $x$ 's together into one weighted sum that we hope will capture all the information they provide about  $y$ .

We then turn the combination into a probability by applying  $F$ .

Inference as in the  $p = 1$  case discussed previously except now our likelihood will depend on  $(\beta_0, \beta_1, \dots, \beta_p)$  instead of just  $(\beta_0, \beta_1)$ .

## The Default Data, More than One x

Here is the logistic regression output using all three x's in the data set: balance, income, and student.

student is coded up as a factor, so R automatically turns it into a dummy.

Call:

```
glm(formula = default ~ balance + student + income, family = binomial,
     data = Default)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.4691	-0.1418	-0.0557	-0.0203	3.7383

Coefficients:

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

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2920.6 on 9999 degrees of freedom

Residual deviance: 1571.5 on 9996 degrees of freedom

AIC: 1579.5

Number of Fisher Scoring iterations: 8

Everything is analogous to when we had one  $x$ .

The estimates are MLE.

Confidence intervals are estimate  $\pm$  2 standard errors.

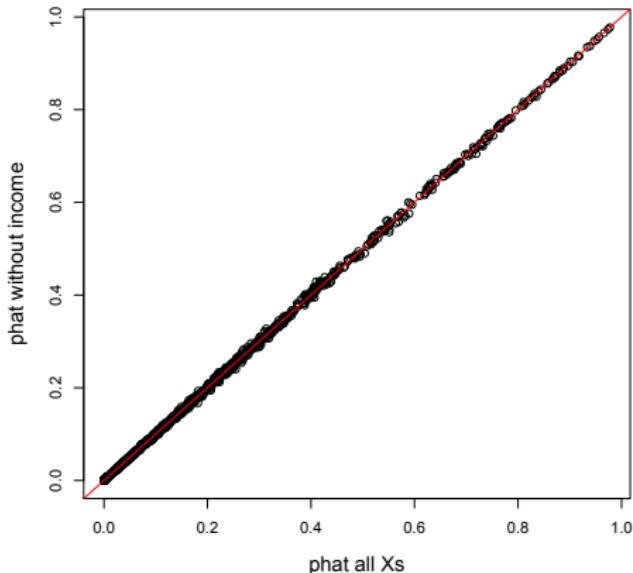
Z-stats are  $(\text{estimate}-\text{proposed})/\text{se}$ .

To test whether the coefficient for income is 0, we have  $z = (3.033-0)/8.203 = .37$ , so we fail to reject.

The p-value is  $2*P(Z < -.37) = 2*pnorm(-.37) = 0.7113825$ .

So, the output suggests we may not need income.

Here is a plot of the fitted probabilities with and without income in the model.



We get almost the same probabilities, so, as a practical matter, income does not change the fit.

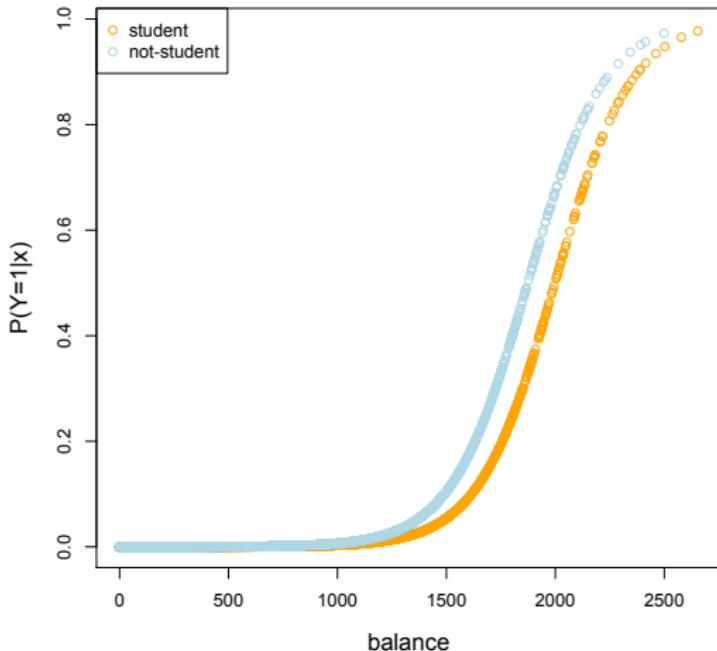
Here is the output using balance and student.

```
Call:  
glm(formula = default ~ balance + student, family = binomial,  
    data = Default)  
  
Deviance Residuals:  
    Min      1Q  Median      3Q     Max  
-2.4578 -0.1422 -0.0559 -0.0203  3.7435  
  
Coefficients:  
            Estimate Std. Error z value Pr(>|z|)  
(Intercept) -1.075e+01  3.692e-01 -29.116 < 2e-16 ***  
balance       5.738e-03  2.318e-04   24.750 < 2e-16 ***  
studentYes   -7.149e-01  1.475e-01  -4.846 1.26e-06 ***  
---  
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 2920.6 on 9999 degrees of freedom  
Residual deviance: 1571.7 on 9997 degrees of freedom  
AIC: 1577.7  
  
Number of Fisher Scoring iterations: 8
```

With just balance and student in the model, we can plot  $P(Y = 1 | x)$  vs.  $x$ .

The orange points are for the students and the blue are for the non-students.

In both cases the probability of default increases with the balance, but at any fixed balance, a student is less likely to default.



## Confounding Example:

The ISLR book notes a nice example of “confounding” in the Default data.

Suppose we do a logistic regression using only student.

Here the coefficient for the student dummy is positive, suggesting that a student is more likely to default.

But, in the multiple logistic regression, the coefficient for student was -.7 and we saw the a student was less likely to default at any fixed level of balance.

```
Call:  
glm(formula = default ~ student, family = binomial, data = Default)  
  
Deviance Residuals:  
    Min      1Q  Median      3Q     Max  
-0.2970 -0.2970 -0.2434 -0.2434  2.6585  
  
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  
  
(Dispersion parameter for binomial family taken to be 1)  
  
Null deviance: 2920.6  on 9999  degrees of freedom  
Residual deviance: 2908.7  on 9998  degrees of freedom  
AIC: 2912.7  
  
Number of Fisher Scoring iterations: 6
```

How can this be?

This is the sort of thing where our intuition from linear multiple regression can carry over to logistic regression. Since both methods start by mapping a  $p$  dimensional  $x$  down to just one number, they have some basic features in common. That is a nice thing about using logistic regression.

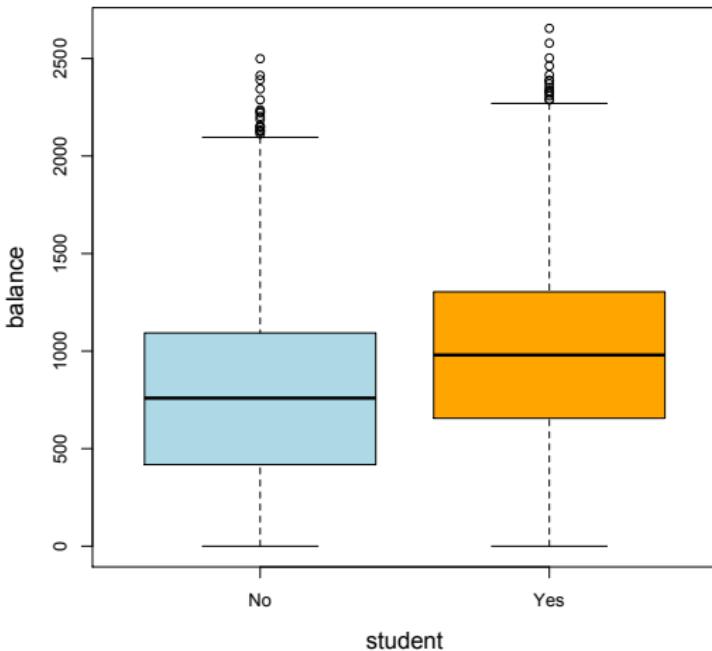
We know that when  $x$ 's are correlated the coefficients for old  $x$ 's can change when we add new  $x$ 's to a model.

Are balance and student “correlated”?

Here is a plot of balance vs. student.  
We can see they are related.

If all you know is that a credit card holder is a student, then (in the background) they are more likely to have a larger balance and hence more likely to default.

But, if you know the balance, a student is less likely to default than a non-student.



## 7. AIC and BIC in Logistic Regression

We have reported logistic regression results for a variety of choices for  $x$ .

- ▶ balance:  
Residual deviance: 1596.5, AIC: 1600.5
- ▶ balance + student + income:  
Residual deviance: 1571.5, AIC: 1579.5
- ▶ balance + student:  
Residual deviance: 1571.7, AIC: 1577.7
- ▶ student:  
Residual deviance: 2908.7, AIC: 2912.7

A smaller residual deviance indicates a better fit.

*But*, it can only get smaller when you add variables!

The deviance is just -2 times the maximized log likelihood. When you add  $x$  variables the maximized likelihood can only get bigger so the deviance can only get smaller.

If you have more coefficients to optimize over you can only do better since you can set them to zero if you want.

This is analogous to the fact that in linear multiple regression  $R^2$  can only go up when you add  $x$ 's.

AIC is analogous to the BIC and adjusted  $R^2$  in that it penalizes you for adding variables.

Rather than choosing the model with the smallest deviance, some people advocate choosing the model with the smallest AIC value:

$$AIC = -2\log(\hat{L}) + 2(p + 1) = \text{deviance} + 2(p + 1),$$

where  $\hat{L}$  is maximized likelihood and  $p$  is the number of  $x$ 's (we add 1 for the intercept).

The idea is that as you add variables (the model gets more complex), deviance goes down but  $2*(p+1)$  goes up.

The suggestion is to pick the model with the smallest AIC.

## AIC for the Default example:

A parameter (a coefficient) costs 2.

- ▶ balance:

Residual deviance: 1596.5, AIC:  $1600.5 = 1593.5 + 2 * (2)$ .

- ▶ balance + student + income:

Residual deviance: 1571.5, AIC:  $1579.5 = 1571.5 + 2 * (4)$ .

- ▶ balance + student:

Residual deviance: 1571.7, AIC:  $1577.7 = 1571.7 + 2 * (3)$ .

- ▶ student:

Residual deviance: 2908.7, AIC:  $2912.7 = 2908.7 + 2 * (2)$ .

⇒ pick balance+student

## BIC:

BIC is an alternative to AIC, but the penalty is different.

$$BIC = \text{deviance} + \log(n) * (p + 1)$$

$\log(n)$  tends to be bigger than 2, so BIC has a bigger penalty, so it suggest smaller models than AIC.

## BIC for the Default example:

$$\log(10000) = 9.21034.$$

A parameter (a coefficient) costs 9.2.

► balance:

$$1596.5, \text{ BIC: } = 1593.5 + 9.2*(2) = 1611.9.$$

► balance + student + income:

$$\text{BIC: } = 1571.5 + 9.2*(4) = 1608.3.$$

► balance + student:

$$\text{BIC: } = 1571.7 + 9.2*(3) = 1599.3.$$

► student:

$$\text{BIC: } = 2908.7 + 9.2*(2) = 2927.1.$$

⇒ pick balance+student

Which is better, AIC or BIC??

*nobody knows.*

R prints out AIC, which suggests you might want to use it, but a lot of people like the fact that BIC suggests simpler models.

A lot of academic papers report both AIC and BIC and if they pick the same model are happy with that. Lame.

*Checking the out of sample performance is safer !!!*

## 8. Making Decisions

Let's consider the simple (*but very important*) case where  $Y$  is 0 or 1.

If  $p(x)$  is our estimate of  $P(Y = 1 | x)$  the obvious thing to do is classify (predict)  $Y = 1$  if  $p(x) > .5$ . We can then look at the missclassification rate.

*Sometimes we need to consider the consequences of our actions more carefully!*

The simple rule “guess  $Y = 1$  if  $p(x) > .5$ ” may not be appropriate.

## Targeted Marketing:

You are trying to decide whether or not to mail a promotion to a household.

$Y = 1$  if they buy something (when mailed the promotion)  
and 0 if they don't.

$x$ : stuff about the household

- ▶ demographics
- ▶ past purchase history.

$p(x)$ :

Probability a household like  $x$  will buy when prompted by promotion.

$p(x)$ : Probability a household like  $x$  will buy when prompted by promotion.

*Typically all  $p(x)$  are much less than .5!!!*

It does not make sense to predict that nobody will buy and consequently send out zero promotions.

In this application,  $p(x) = .1$  is *huge* since there is a “real” chance they will respond and then spend an amount which is large relative to the cost of the mailing.

## Minimizing Expected Loss

In our Target Marketing example you want to mail a promotion if:

- ▶  $p(x)$ , the probability of a purchase is big.
- ▶  $A$ , the average amount spent is big.
- ▶  $c$ , the cost of a mailing is small.

Let's formally go through how you make the decision to send a mailing given these inputs.

Of course, in a given problem there may be other factors to consider, but this simple framework allows us to explore the idea of using probability to make a decision.

## General Framework:

In general let's assume we are uncertain about  $Y$ , which will be 1 or 0.

We have to decide to do something, or not:  $d = 1$  means you do it, 0 means you do not.

There are then four different possible outcomes depending on the random  $Y$  and our decision.

Let  $L(y, d)$  be the loss if  $Y = y$  and you decide  $d$ , where  $y$  and  $d$  are 0 or 1.

Let  $L(y, d)$  be the loss if  $Y = y$  and you decide  $d$ , where  $y$  and  $d$  are 0 or 1.

*Make the decision which gives the smallest expected loss!*

if  $d = 0$ , expected loss is:

		d	
		0	1
y	0	$L(0, 0)$	$L(0, 1)$
	1	$L(1, 0)$	$L(1, 1)$

$$\begin{aligned}E(L(Y, 0)) &= \\(1 - p(x)) L(0, 0) + p(x) L(1, 0)\end{aligned}$$

if  $d = 1$ , expected loss is:

$$\begin{aligned}E(L(Y, 1)) &= \\(1 - p(x)) L(0, 1) + p(x) L(1, 1)\end{aligned}$$

*Our optimal decision is:*

$d(x) = 1$  if:

$$(1 - p(x)) L(0, 1) + p(x) L(1, 1) < (1 - p(x)) L(0, 0) + p(x) L(1, 0).$$

and 0 otherwise.

## Target Marketing:

		$d$		if $d = 0$ , expected loss is: $p(x) A$
	0	0	$c$	
$y$	1	$A$	$c$	if $d = 1$ , expected loss is: $c$

Optimal decision:

$$d(x) = \begin{cases} 0 & p(x) A < c \\ 1 & c < p(x) A \end{cases}$$

Pretty obvious: mail if expected benefits are greater than the cost.

Given our decision rule, we can express our loss in terms of  $Y$  and  $x$ :

$$L(y, x) = L(y, d(x)).$$

Target Marketing:

$$L(y, x) = L(y, d(x)) = \begin{cases} c & p(x) > \frac{c}{A} \\ y A & p(x) < \frac{c}{A} \end{cases}$$

*Clearly, the rule classify  $Y = 1$  if  $p(x) > .5$  is not relevant!*

## Another Credit Example

Assume we were after a predictive model to determine who is a good customer, i.e., who has a high probability of paying back the loan. (What tools did we use here?)

So, when a new customer walks in the bank and asks for a loan we are able to predict the probability of payback given a set of customers characteristics (features)... this could be  
 $\hat{p} = 0.3, \hat{p} = 0.5, \hat{p} = 0.8\dots$

We need to decide on a cutoff to extend or not the loan

Why not just choosing 0.5, i.e., if a costumer is more likely than not to pay, give him the money?

## Another Credit Example

Well, we might have different **costs** associated with a default and not given a loan to a potentially good customer. In classification lingo, a **false-positive** might cost more than a **false-negative**!

For example, imagine we have the following cost table for this situation:

	Loan	No-loan
$(1 - p)$	500	0
$p$	0	100

The expected cost under “loan” is  $500(1 - p)$

The expected cost under “no-loan” is  $100p$

The costs are equal when  $p = 5/6\dots$  therefore we only loan money to customer with a  $\hat{p} > 5/6$ !

## 9. Out of Sample

We have seen how loss considerations help us decide on a good decision in the face of an uncertain binary  $Y$ .

A good  $p(x)$  is one where our average *out of sample* loss is smallest!!

We want

$$E(L(Y, d(X)))$$

to be small where the expectation  $E$  averages over future  $(Y, X)$  pairs.

Given out-of-sample observations  $(X_i^0, Y_i^o)$ ,  $i = 1, 2, \dots, m$  we estimate the out-of-sample expected loss with

$$\frac{1}{m} \sum_{i=1}^m L(Y_i^o, d(X_i^o)).$$

Note:

We do the same thing for numeric  $Y$ !!!

Our notation for numeric  $Y$  is to predict  $Y$  given  $X = x$  with  $d(x) = \widehat{f(x)}$ .

Our loss is then

$$L(y, x) = L(y, d(x)) = (y - \widehat{f(x)})^2.$$

We estimate the out of sample expected loss with:

$$MSE^o = \frac{1}{m} \sum_{i=1}^m (Y_i^o - \widehat{f(X_i^o)})^2$$

We often then take the square root to put the units on the original  $Y$  scale giving  $RMSE^o$ .

## Deviance As Out of Sample Loss

In our target marketing example, we thought about the actual consequences of our decision in our applied context to come up with our loss. *This is the right thing to do!*

Often however, we are not sure how we want to use our predictor, or are just too plain busy to think about it carefully.

In this case, a generic loss is useful.

For numeric  $Y$   $MSE$  (or  $RMSE$ ) are the most commonly used generic loss functions.

For classification problems with categorical  $Y$ , we often use the **missclassification rate** (see Section 1) for the decision rule which picks the most likely outcome.

In the binary  $Y$  case, our targeting marketing example illustrates the point that a lot of time we have important fit and  $p(x) < .5$  for almost all  $x$ .

The ROC curve and the lift look at the performance of the rule  $p(x) < s$  as  $s$  is varied from 0 to 1.

Alternatively, the *deviance* is used as a generic loss function for classification problems.

## Deviance Loss (aka log-loss):

Let

$$\hat{p}_y = \widehat{P(y | x)}.$$

Given out-of-sample observations  $(X_i^0, Y_i^o)$ ,  $i = 1, 2, \dots, m$  we estimate the out-of-sample expected loss with

$$\sum_{i=1}^m -2 \log(\hat{p}_{Y_i^o}).$$

As discussed in section 5, this is just the sum of our losses when  $Y_i^o$  happens and we had assigned that outcome probability  $\hat{p}_{Y_i^o}$ .

## Deviance Loss (aka log-loss):

Why deviance/log-loss?

Statistical theory of maximum likelihood (for logistic regression)

Deviance/Log-loss is a *proper loss*: It encourages *sharpness* (predicted probabilities away from 0.5) while maintaining *calibration* (for all cases where the predicted probability is  $p$ , 100( $p$ )% are 1).

Calibration is important for accurately estimating conditional expected values (making decisions)!

## Default Data

Let's go back to the default data and use out of sample performance to do variable selection.

Let's do a simple train/test split.

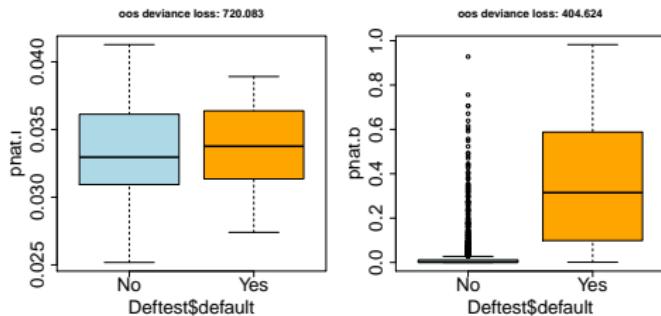
- ▶ train: randomly pick 75%.
- ▶ test: the remaining 25%.

For each choice of variables, fit on train  
and then plot  $y = \text{default}$  vs  $\widehat{p(x)}$  for the test data.

(1,1):

$x$  is income.

out of sample deviance loss:  
720.083.



(1,2):

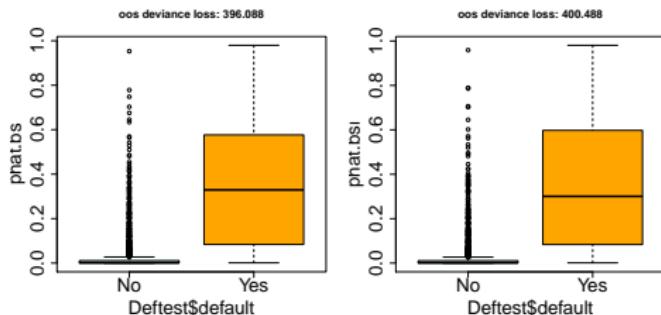
$x$  is balance.

out of sample deviance loss:  
404.624.

(2,1):

$x$  is balance+student.

out of sample deviance loss:  
396.088.



(2,1):

$x$  is balance+student+income.

out of sample deviance loss:  
400.488.

*Looks like just using  $x=\text{balance}$  works pretty good!!*

Let's put the train/test split into a loop (like cross-validation) to make sure our result is not sensitive to our random split.

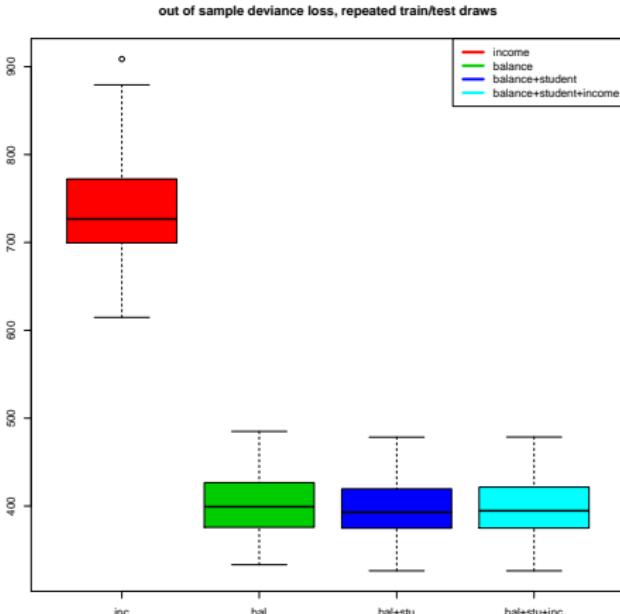
100 times we randomly split into train/test (75/25).

Each time we compute the out-of-sample deviance loss for each choice of variables.

Income alone is terrible.

No “significant” difference between the other three.

Just  $x=balance$   
still looks good!



## 10. Lift and ROC

Lift and ROC are two popular methods for assessing the quality of a classifier for a binary  $y$ .

Lift is particularly popular in Marketing.

ROC stands for the incomprehensible term “receiver operator characteristics”.

Both look at missclassification rates for various values of  $s$  using the rule: classify  $Y = 1$  if  $p(x) > s$ .

Let's look at our trusty default data again.

Of course, it is interesting to think about what your actual decision problem is in the default data context!!!

Kick them out if  $p(x) > .2$ ?

Send them a nasty letter if  $.1 < p(x) < .2$ ?

Again, lift and ROC look at the performance of “ $Y = 1$  if  $p(x) > s$ ” for a variety of  $s$  values. It is not what you would actually do, but it gives a sense of how  $p(x)$  is doing.

To start with, suppose  $s = .5$ .

We use all the data and fit a logit using  $x=balance$ .

You could (should) of course, do this out of sample.

For each observation:

$\hat{y} = 1$  if  $p(x) > .5$  and 0 otherwise.

Here is the table relating  $y$  to  $yhat$ :

$y$	0	1
$yhat$	9625	233
0	42	100

This table is called the  
*confusion matrix*.

As the book notes, a *bewildering*  
number of terms have been invented  
to talk about this simple  $2 \times 2$  table!

Counts on the diagonal are success, while the off-diagonal  
represent the two kinds of failures you can make.

ROC and lift summarize the situation.

	y	
yhat	0	1
0	9625	233
1	42	100

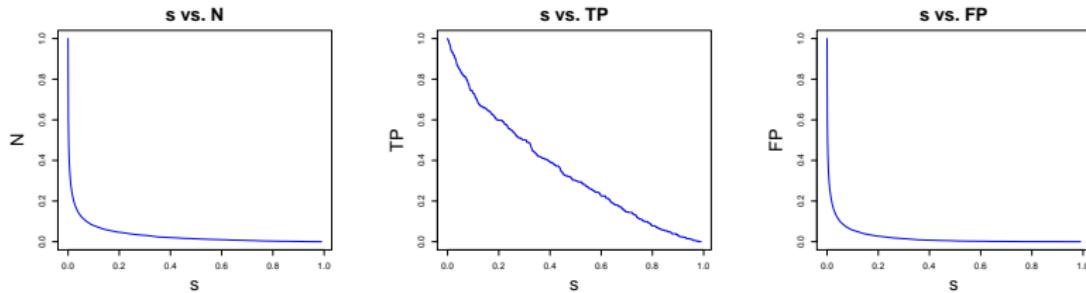
ROC looks at:

- ▶ TP (true positive), %  $y=1$  correctly classified:  
 $100/(100+233) = .3.$
- ▶ FP (false positive), %  $y=0$  incorrectly classified:  
 $42/(9625+42) = 0.0043.$

Lift looks at:

- ▶ TP (true positive), %  $y=1$  correctly classified: .3.
- ▶ N, % classified as 1:  $(100+42)/10000=.0142.$

We then just compute these values for  $s$  between 0 and 1 and plot them.

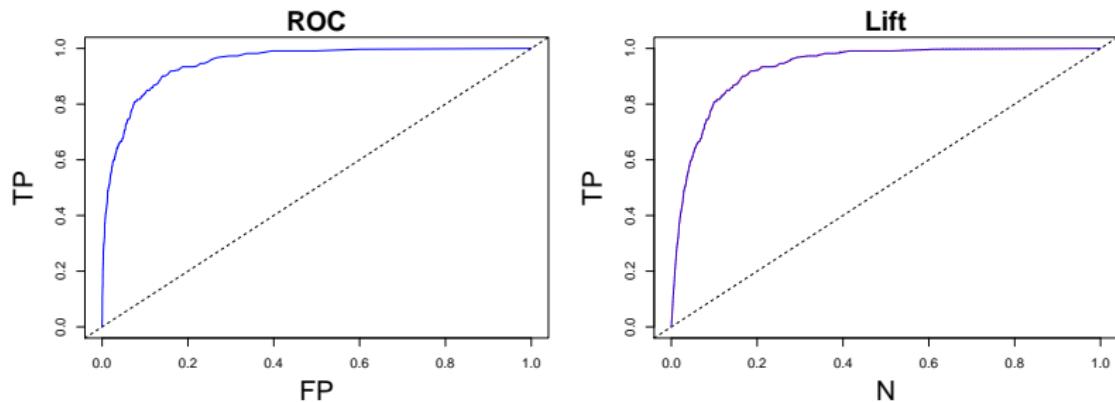


$$\hat{y} = 1 \Leftrightarrow p(x) > s.$$

- ▶ TP (true positive), %  $y=1$  correctly classified
- ▶ FP:(false positive), %  $y=0$  incorrectly classified
- ▶ N, % classified as 1.

All three quantities go from 1 to 0, as  $s$  goes from 0 to 1.

ROC plots FP vs. TP, and lift plots N vs. TP.



Note that as you go from left to right in these plots,  $s$  is decreasing.

The line is drawn at " $y=x$ ".

It represents the performance you would get if  $Y$  was independent of  $X$ .

## More on Lift:

There is another simple way to think about the lift curve without referring to the cutoff  $s$ . Again let's use a Marketing problem to motivate things.

Suppose you have a budget that allows you to mail out to  $N$  percent of the potential customers you have on file.

For customer with information  $x$ ,  $p(x)$  tells you how likely they are to respond.

Given budget  $N$ , you mail out to the  $N$  percent of potential customers on file that have the large values of  $p(x)$ .

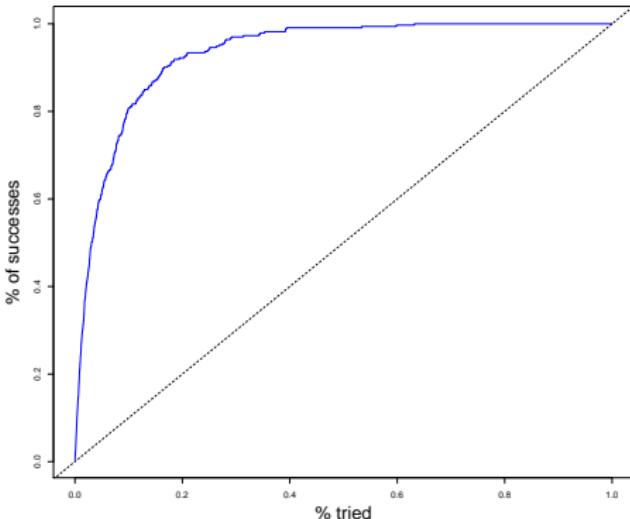
For each  $N$  (% of the ones you try) compute

$$\frac{\text{number of } (y=1) \text{ from } N\%}{\text{number of } (y=1) 100\%} = \% \text{ of the good ones you got.}$$

(which is just TP again).

Here is the lift again, but just relabeled:  
how many you tried vs. how many of the good ones you got.

If you were just guessing, you would get  $x\%$  of the good ones (on average) if you tried  $x\%$  of the cases. This is captured by the straight line.



How much the lift curve is above the straight line gives you a feeling for how good  $p(x)$  is.

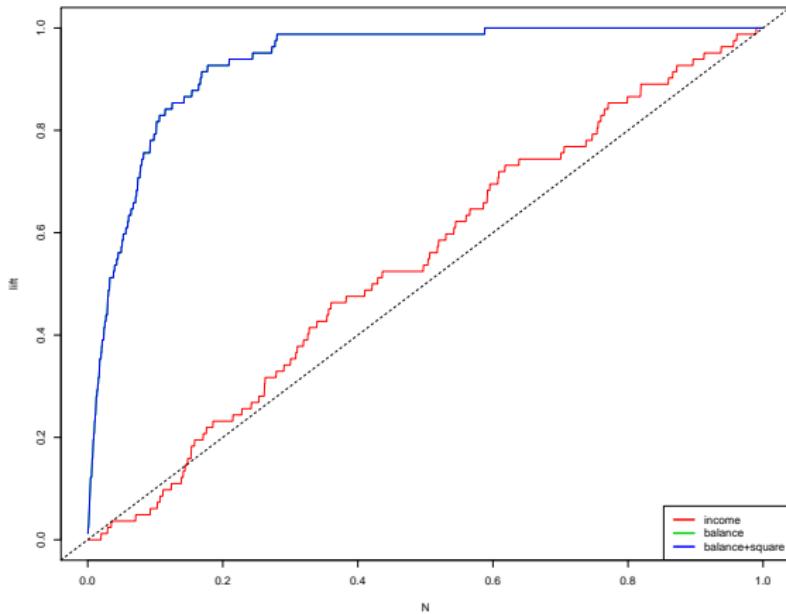
Let's use lift to compare some classifiers on the default data.

Let's try logits with:

- (i)  $x = \text{income}$ .
- (ii)  $x = \text{balance}$ .
- (iii)  $x = \text{balance} + \text{balance squared}$ .

Each one will give us a  $p(x)$  and will compare them based on their out-of-sample lift.

The lift for income is so bad, it is very close to the line.



The lift for balance + balance squared is right on top of the one for just using balance.

It did not hurt to throw in the square, but it did not help.

## 11. Bayes Theorem and Classification

We noted that you can think of logistic regression as a parametric model for

$$P(Y = y | X = x)$$

the *conditional distribution of Y given X = x*.

A number of classification techniques work by specifying the marginal distribution of  $Y$ , the conditional distribution of  $X$  given  $Y$  and then using Bayes Theorem to compute the conditional distribution of  $Y$  given  $X$ .

Conceptually, this is a very nice approach.

But it is tricky in that a lot of probability distributions have to be specified. In particular, you have to specify the possibly high-dimension distribution of  $X$  given  $Y$ .

Up to now we have been intuitive in our use of probability.  
Let's quickly review the basic definitions and Bayes Theorem.

## 11.1. Quick Basic Probability Review

Suppose  $X$  and  $Y$  are discrete random variables.

This means we can list out the possible values.

For example, suppose  $X$  can be 1,2, or 3, and  $Y$  can be 0 or 1.

Then we specify the joint distribution of  $(X, Y)$  by listing out all the possible pairs and assigning a probability to each pair:

For each possible  $(x, y)$  pair  $p(x, y)$  is the probability that  $X$  turns out to be  $x$  and  $Y$  turns out to be  $y$ .

$$p(x, y) = \Pr(X = x, Y = y).$$

**Note:**  $X$  is the random variable.  $x$  is a possible value  $X$  could turn out to be.

$x$	$y$	$p(x, y)$
1	0	.894
2	0	.065
3	0	.008
1	1	.006
2	1	.014
3	1	.013

We can also arrange the probabilities in a nice two-way table.

columns:

indexed by  $y$  values

rows:

indexed by  $x$  values.

		y	
		0	1
		1	.894 .006
x	2	.065	.014
	3	.008	.013

Where did these numbers come from?

These numbers are an estimate of the joint distribution of default and a *discretized* version of balance from the Default data.

$Y$  is just 1 (instead of Yes) for a default,  
and 0 otherwise (instead of No).

To discretize balance we let  $X$  be

- ▶ 1 if balance  $\leq 1473$ .
- ▶ 2 if  $1473 < \text{balance} \leq 1857$ .
- ▶ 3 if  $1857 < \text{balance}$ .

This gives the simple two-way table of counts:

def	bal	0	1
1	8940	64	
2	651	136	
3	76	133	

With corresponding percentages (divide by 10,000):

def	bal	0	1
1	0.894	0.006	
2	0.065	0.014	
3	0.008	0.013	

Normally, we might use names like  $D$  and  $B$  for our two variables, but since we want to think about the ideas in general, let's stick with  $Y$  and  $X$ .

## Joint Probabilities:

$p(x, y) = P(X = x, Y = y)$ , the probability that  $X$  turns out to be  $x$  *and*  $Y$  turns out to be  $y$  is called the *joint probability*.

The complete set of joint probabilities specifies the *joint distribution* of  $X$  and  $Y$ .

## Marginal Probabilities:

Given the joint distribution, we can compute the *marginal probabilities*  $p(x) = P(X = x)$  or  $P(Y = y)$ .

$$p(x) = \sum_y p(x, y), \quad p(y) = \sum_x p(x, y).$$

Computing marginal probabilities from a joint:

$$P(Y=1) =$$

$.006 + .014 + .013 = .033$

$$P(X=3) =$$

$0.008 + 0.013 = 0.021$

$x$	$y$	$p(x, y)$
1	0	.894
2	0	.065
3	0	.008
1	1	.006
2	1	.014
3	1	.013

$x$	$y$	$p(x, y)$
1	0	.894
2	0	.065
3	0	.008
1	1	.006
2	1	.014
3	1	.013

## Conditional Probabilities:

$P(Y = y | X = x)$  is the probability  $Y$  turns out to be  $y$  *given* you found out that  $X$  turned out to be  $x$ .

This fundamental concept is how we quantify the idea of updating our beliefs in the light of new information.

$$P(Y = y | X = x) = \frac{p(x, y)}{p(x)}.$$

The fraction of times you get  $x$  and  $y$  out of the times you get  $x$ .

or,

$$p(x, y) = p(x) p(y | x).$$

The chance of getting  $(x, y)$  is the fraction of times you get  $x$  times the fraction of those times you get  $y$ .

$$P(Y = 1 \mid X = 3)$$

$$= \frac{p(3,1)}{p(3)}$$

$$= \frac{.013}{.008+.013}$$

$$= \frac{.013}{.021} = .62.$$

$$P(X = 3 \mid Y = 1)$$

$$= \frac{p(3,1)}{p(1)}$$

$$= \frac{.013}{.006+.014+.013}$$

$$= \frac{.013}{.033} = .394.$$

$x$	$y$	$p(x, y)$
1	0	.894
2	0	.065
3	0	.008
1	1	.006
2	1	.014
3	1	.013

$x$	$y$	$p(x, y)$
1	0	.894
2	0	.065
3	0	.008
1	1	.006
2	1	.014
3	1	.013

You just renormalize the relevant probabilities given the information!!

Compare:

$$P(Y=1) = .033$$

$$P(X=3) = .021.$$

## 11.2. Conditional Probability and Classification

Clearly, we can use  $P(Y = y | X = x)$  to classify given a new value of  $x$ .

The most obvious thing to do is predict the  $y$  that has the highest probability.

Given  $x = x_f$ , we can predict  $Y$  to be  $y_f$  where

$$P(Y = y_f | X = x_f) = \max_y P(Y = y | X = x_f).$$

Remember, we are assuming there is just a small number of possible  $y$  so you just have to look at see which one is biggest.

For our example with default ( $Y$ ) and discretized balance ( $X$ ) our joint is

def	bal	0	1
1	0.894	0.006	
2	0.065	0.014	
3	0.008	0.013	

If we simply divide each row by its sum we get the conditional of  $Y=\text{default}$  given  $X=\text{balance}$ .

def	bal	0	1	So, not surprisingly,
1	0.993	0.007		if we use the max prob rule,
2	0.823	0.177		you are classified (predicted)
3	0.381	0.619		as a potential defaulter if balance=3.

### Note:

If there are only two possible outcomes for  $Y$ , we are just picking the one with  $P(Y = y | x) > .5$ .

But, it is a nice feature of this way of thinking that it works pretty much the same if  $Y$  is multinomial (more than two possible outcomes) rather than just binomial (two outcomes).

### Note:

Since the probabilities have to add up to one, the chance of being wrong is just

$$1 - \max_y P(Y = y | X = x_f).$$

So, in our previous example, the error probabilities are .007, .177, and .381 for  $x=\text{default} = 1, 2, 3$  respectively.

## 11.3. Bayes Theorem

In the previous section we saw that if  $Y$  is discrete, and we have the joint distribution of  $(X, Y)$  we can “classify”  $y$  by computing  $P(Y = y | X = x)$  for all possible  $y$ .

Note that a nice feature of this approach is that it naturally handles the case where  $Y$  can take on more than two values.

Logistic regression assumes two categories for  $Y$ .

There is a *multinomial* version of logistic regression but it is more complex.

When we use Bayes Theorem for classification we again compute  $P(Y = y | X = x)$ .

However we assume that we specify the joint distribution by specifying:

- ▶ the marginal distribution of  $Y$ .
- ▶ the conditional distribution of  $X$  given  $Y$ .

That is, we have to specify:

$$p(y) \text{ and } p(x | y).$$

“Bayes Theorem” simply says that if I give you  $p(y)$  and  $p(x | y)$ , you can compute  $p(y | x)$ .

This is obvious since we know  $p(x, y) = p(y)p(x | y)$  and if we have the joint we can compute either conditional.

To follow the notation in the book let's write  $k$  for  $k = 1, 2, \dots, K$  for the possible values of  $Y$  instead of  $y$ . We then want  $P(Y = k | X = x)$ .

Bayes Theorem:

$$P(Y = k | X = x) = \frac{p(x, k)}{p(x)} = \frac{p(x, k)}{\sum_{l=1}^K p(x, l)} = \frac{p(Y = k)p(x | k)}{\sum_{l=1}^K p(Y = l)p(x | l)}.$$

$$P(Y = k \mid X = x) = \frac{p(Y = k)p(x \mid k)}{\sum_{l=1}^K p(Y = l)p(x \mid l)}.$$

To further match up the notation that of the book, let

$$P(Y = k) = \pi_k, \text{ and } p(x \mid k) = f_k(x).$$

We then have:

$$P(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}.$$

(see equation 4.10 in the book)

## Default Example:

In our example, the idea is that you start off knowing

### (I) The Marginal of $Y$ .

$$\pi_1 = P(Y = 1) = .9667, \quad \pi_2 = P(Y = 2) = .0333.$$

(now  $k = 2$ , annoyingly, means a default and  $k = 1$  means no default.)

### (II) the conditional distribution of $X$ for each $y$

def	1	2
bal	1	2
1	0.925	0.182
2	0.067	0.424
3	0.008	0.394

Take the table giving the joint  $p(x, y)$  and renormalize the columns so that they add up to one.

Column 1 is  $P(X = x | Y = 1) = f_1(x)$ .

Column 2 is  $P(X = x | Y = 2) = f_2(x)$ .

So, for example,

$$P(X = 2 | Y = 1) = f_1(2) = .067.$$

So, suppose you know  $X = 3$ , how to you classify  $Y$ ?

$$\begin{aligned} P(Y = 2 \mid X = 3) &= \frac{\pi_2 f_2(3)}{\pi_1 f_1(3) + \pi_2 f_2(3)} \\ &= \frac{.0333 * .394}{.9667 * .008 + .0333 * .394} \\ &= \frac{.013}{.008 + .013} \\ &= .62. \end{aligned}$$

as before.

Even though defaults ( $Y = 2$ ) are unlikely, *after seeing  $X=3$ ,  $Y=2$  is likely because seeing  $X=3$  is much more likely if  $Y=2$  (.394) than if  $Y=1$  (.008).*

Note:

The  $\pi_k$  are called the *prior* class probabilities.

This is how likely you think  $Y = k$  is *before* you see  $X = x$ .

Note:

A succinct way to state Bayes Theorem is

$$P(Y = k | X = x) \propto \pi_k f_k(x).$$

where  $\propto$  means “proportional to”.

$P(Y = k | X = x)$  is called the *posterior* probability that  $Y = k$ .

This is how likely you think  $Y = k$  is *after* you see  $X = x$ .

$$P(Y = k \mid X = x) \propto \pi_k f_k(x)$$

$\pi_k$ : how likely case  $k$  is for  $Y$  before you see the data  $x$ .

$f_k(x)$ : how likely the data  $x$  is, given  $Y$  is in case  $k$ .

Here,  $f_k(x)$  is our *likelihood*, it tells us how likely what we saw is for different values of  $k$ .

Basic Intuition: If you see something that was likely to happen if  $Y = k$ , maybe  $Y = k$  !!

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

Bayes Theorem beautifully combines our prior information with the information in the data.

Let's redo the example using the proportional form of the formula:

$$\begin{aligned} P(Y = 1 | X = 3) &\propto \pi_1 f_1(3) \\ &= .9667 * .008 \\ &= 0.0077336. \end{aligned}$$

$$\begin{aligned} P(Y = 2 | X = 3) &\propto \pi_2 f_2(3) \\ &= .0333 * .394 \\ &= 0.0131202. \end{aligned}$$

$$P(Y = 2 | X = 3) = \frac{0.0131202}{0.0077336 + 0.0131202} = .629.$$

as before.

**Note:**

There is a lot of theory that basically says Bayes Theorem is the right thing to do.

However this assumes the  $\pi_k$  and  $f_k(x)$  are “right”, and we are almost never sure of this.

## 11.4. Naive Bayes

$$P(Y = k \mid X = x) \propto \pi_k f_k(x)$$

To make this exciting we need to make  $x$  high dimensional!!

Since we are doing classification, we still think of  $Y$  as a discrete random variable so we think of the  $\pi_k$  the same way.

However, now we want to think of  $x$  as possibly containing many variables.

Now  $X$  is a vector of random variables  $X = (X_1, X_2, \dots, X_p)$ .

Our probability laws extend nicely in that we still have

$$p(x, y) = P(X = x, Y = y) = p(y)p(x | y) = p(x)p(y | x)$$

If each  $X_i$  is discrete,

$$p(x) = P(X = x) = P(X_1 = x_1, X_2 = x_2, \dots, X_p = x_p).$$

and,

$$f_k(x) = P(X_1 = x_1, X_2 = x_2, \dots, X_p = x_p | Y = k).$$

*And we still have*

$$P(Y = k | X = x) \propto \pi_k f_k(x)$$

Our problem is now obvious.

In practice, how do you specify

$$f_k(x) = P(X_1 = x_1, X_2 = x_2, \dots, X_p = x_p \mid Y = k).$$

for large  $p$ .

This would involve understanding something about the high dimensional  $x$ .

The naive Bayes solution is to assume that, conditional on  $Y$ , all the  $X_i$  are independent.

Let's do  $p = 2$  first so we can simply see what this implies.  
It is always true that:

$$\begin{aligned}f_k(x) &= f_k(x_1, x_2 \mid Y = k) \\&= P(X_1 = x_1, X_2 = x_2 \mid Y = k) \\&= P(X_1 = x_1 \mid Y = k) P(X_2 = x_2, X_1 = x_1 \mid Y = k).\end{aligned}$$

Naive Bayes then assumes that

$$P(X_2 = x_2, X_1 = x_1 \mid Y = k) = P(X_2 = x_2 \mid Y = k).$$

(given  $Y$ ,  $X_1$  has no information about  $X_2$ )

So,

$$f_k(x) = f_k^1(x_1) f_k^2(x_2), \quad f_k^i(x_i) = P(X_i = x_i \mid Y = k).$$

## Naive Bayes:

For general  $p$  we have:

$$f_k(x) = \prod_{i=1}^p f_k^i(x_i)$$

and, as before,

$$P(Y = k \mid X = x) \propto \pi_k f_k(x)$$

## Default Example:

Will will do  $p = 2$  by using student status in addition to balance.  
Let's think of balance (still discretized) as  $x_1$  and student status as  $x_2$ . Student status is a binary variable.

The simple part of Naive Bayes, is that we look at the components of  $x$  one at a time.

So, we still use:

def		
bal	1	2
1	0.925	0.182
2	0.067	0.424
3	0.008	0.394

$P(X_1 = 2 | Y = 1) = f_2^1(3) = .394.$

Here is the joint of  $(X_2, Y) = (\text{student}, \text{default})$ .

def	student	1	2
No	0.685	0.021	
Yes	0.282	0.013	

Here are the conditionals of student, given  $Y = 1$  or  $2$ .

def	student	1	2
No	0.708	0.618	
Yes	0.292	0.382	

Thinking of No and Yes as 1 or 2, we have, for example:  
 $f_1^2(2) = P(X_2 = 2 | Y = 1) = .292$ .

OK, we ready to go, our information is:

(I) The Marginal of  $Y=\text{default}$ .

$$\pi_1 = P(Y = 1) = .9667, \quad \pi_2 = P(Y = 2) = .0333.$$

(II) The conditional distributions of  $X_1=\text{balance}$  and  $X_2=\text{student}$

bal	def	1	2
1	0.925	0.182	
2	0.067	0.424	
3	0.008	0.394	

student	def	1	2
No	0.708	0.618	
Yes	0.292	0.382	

So, suppose you know  $X_1 = 3$  and  $X_2 = 2$ , how to you classify  $Y$ ?

$$\begin{aligned} P(Y = 1 \mid X_1 = 3, X_2 = 2) &\propto \pi_1 f_1^1(3) f_1^2(2) \\ &= .9667 * .008 * .292. \\ &= 0.002258211. \end{aligned}$$

$$\begin{aligned} P(Y = 2 \mid X_1 = 3, X_2 = 2) &\propto \pi_2 f_2^1(3) f_2^2(2) \\ &= .0333 * .394 * .382 \\ &= 0.005011916. \end{aligned}$$

$$P(Y = 2 \mid X_1 = 3, X_2 = 2) = \frac{0.005011916}{0.002258211 + 0.005011916} = .689.$$

**Note:**

Just knowing  $X_1 = 3$  (high balance) we got  $P(Y = 2 | \text{info}) = .62$  (probability of a default is .62).

Knowing  $X_1 = 3$  (high balance) *and*  $X_2 = 2$  (a student) we got  $P(Y = 2 | \text{info}) = .69$  (probability of a default is .69.)

Knowing the student status changed things quite a bit.

**Note:**

If you compare the calculation of  $\pi_k f_k(x)$  with just  $X_1$  versus the one with  $X_1$  and  $X_2$ , we see that we just multiplied in an additional term for  $P(X_2 = 2 \mid Y = k) = f_k^2(2)$ .

With more  $x$ 's you would just keep multiplying in an additional contribution for each  $X_j, j = 1, 2, \dots, p!!!$

The “scales” beautifully, in that the computation is linear in  $p$ .

*But*

(i)

You do have to think carefully about each  $X_j$  to come up with its conditional given  $Y$ .

(ii)

The word “naive” in the name comes from the assumption that the  $X$ ’s are independent given  $Y$ .

We know balance and student are not independent, but are they independent given the default status?

*However* the folklore is that Naive Bayes works surprisingly well!!