Classification

James Scott

ECO 395M: Data Mining and Statistical Learning

Outline

1. Linear probability model
2. Logistic regression: the basics
3. Interpreting a logit model
4. Estimating a logit model: MLE
5. Error rates and ROC curves
6. KNN for classification
7. The multinomial logit model, a.k.a "softmax"
8. Evaluating a classifier: likelihood and deviance
9. Naive Bayes classification

Classification

In classification, the target variable y is membership in a category.

- occupation: butcher, baker, candlestick maker...
- consumer choices: Hyundai, Toyota, Ford...
- college major: economics, mathematics, literature...
- tumor type: malignant or benign
- political slant of a new article: R or D

Each observation consists of:

- an observed class $y_i \in \{1, \ldots, M\}$
- a vector of features x_i.

The classification problem: given new $x^\star$, predict $y^\star$ (or provide $P(y^\star = k)$ for each class k).

Linear probability model

We'll start with binary classification (where we label the outcomes y as 0 for no or 1 for yes, appropriately defined in context).

Recall the basic form of a supervised learning problem:

$$E(y \mid x) = f(x)$$

If y is binary (0/1), then:

$$\begin{aligned} E(y \mid x) &= 0 \cdot P(y = 0 \mid x) + 1 \cdot P(y = 1 \mid x) \\ &= P(y = 1 \mid x) \end{aligned}$$

Conclusion: the expectation of a binary outcome is the probability that the outcome is 1.

Linear probability model

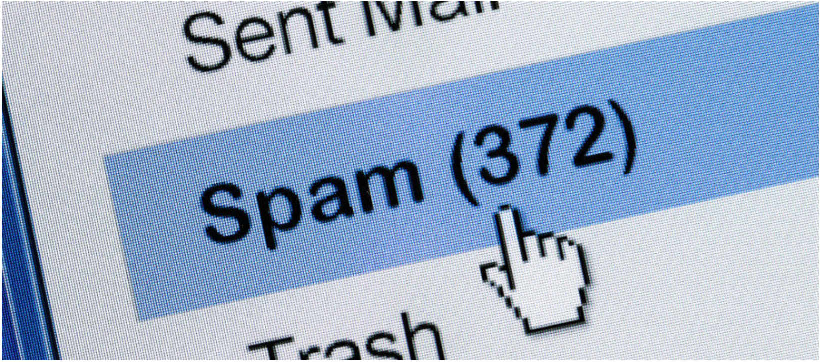
Suppose we choose $f(x)$ to be a linear function of the features x_i:

$$\begin{aligned} P(y = 1 \mid x) &= E(y \mid x) \\ &= x \cdot \beta \\ &= \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \end{aligned}$$

This is called the *linear probability model*: the probability of a "yes" outcome $(y = 1)$ is linear in x_i. To fit this, we:

- Code our outcomes y_i as a dummy variable.
- Throw them into a linear regression model and pretend they're numbers!
- The resulting model predictions give us fitted probabilities.

LPM: spam classification



Let's consider a simple spam classification problem:

- `spamfit.csv`: 3000 e-mails (40% spam) with 9 features.
- `spamtest.csv`: 601 testing e-mails for assessing performance.

LPM: spam classification

Here are the first few lines of our testing data.

```r
spamfit = read.csv('../data/spamfit.csv')
spamtest = read.csv('../data/spamtest.csv')

# first few lines
head(spamtest, 3)
```

```
##   word.freq.remove word.freq.order word.freq.free word.freq.meeting
## 1                0            0.00           5.35                 0
## 2                0            0.00           0.00                 0
## 3                0            0.31           0.63                 0
##   word.freq.re word.freq.edu char.freq.semicolon char.freq.exclamation
## 1            0             0                   0                 0.357
## 2            0             0                   0                 1.975
## 3            0             0                   0                 0.055
##   capital.run.length.average y
## 1                      1.971 1
## 2                     35.461 1
## 3                      3.509 1
```

LPM: spam classification

Let's build a linear probability using all the available features for P(spam | x) and examine the fitted coefficients:

```
# Recall: the dot (.) says "use all variables not otherwise named"
lm_spam1 = lm(y ~ ., data=spamfit)
coef(lm_spam1) %>% round(3)
```

```
##                 (Intercept)             word.freq.remove
##                       0.281                        0.311
##            word.freq.order               word.freq.free
##                       0.284                        0.097
##          word.freq.meeting                 word.freq.re
##                      -0.059                       -0.039
##              word.freq.edu         char.freq.semicolon
##                      -0.051                       -0.096
##      char.freq.exclamation capital.run.length.average
##                       0.229                        0.001
```

LPM: spam classification

Let's look at our in-sample performance by:

- thresholding our predicted probabilities at 0.5
- calculating the **confusion matrix**, which tabulates predicted class versus actual class.

```
phat_train_spam1 = predict(lm_spam1, spamfit)
yhat_train_spam1 = ifelse(phat_train_spam1 > 0.5, 1, 0)
confusion_in = table(y = spamfit$y, yhat = yhat_train_spam1)
confusion_in
```

```
##    yhat
## y      0    1
##   0 1732   68
##   1  541  659
```

```
sum(diag(confusion_in))/sum(confusion_in)  # in-sample accuracy
```

```
## [1] 0.797
```

LPM: spam classification

Let's do the same with out-of-sample performance:

```r
phat_test_spam1 = predict(lm_spam1, spamtest)
yhat_test_spam1 = ifelse(phat_test_spam1 > 0.5, 1, 0)
confusion_out = table(y = spamtest$y, yhat = yhat_test_spam1)
confusion_out  # confusion matrix
```

```
##    yhat
## y     0   1
##   0 372  13
##   1  98 118
```

```r
sum(diag(confusion_out))/sum(confusion_out)  # out-of-sample accuracy
```

```
## [1] 0.8153078
```

LPM: spam classification

How well is our model doing? To answer this question, it helps to have a baseline.

Note that 60% of the training set isn't spam:

```r
table(spamfit$y)
```

```
##
##    0    1
## 1800 1200
```

Since “not spam" is the most likely outcome, a reasonable baseline or “null model” is one that guesses “not spam” for every test-set instance.

LPM: spam classification

How well does this null model perform on the test set? About 64%, since it gets all the 0's right and 1's wrong:

```
table(spamtest$y)
```

```
##
##   0   1
## 385 216
```

```
385/sum(table(spamtest$y))
```

```
## [1] 0.640599
```

Our linear probability model had an 81.5% out-of-sample accuracy rate. Therefore, compared to the null model:

- Its absolute improvement is $\approx 81.5 - 64.1 = 17.4\%$.
- Its relative improvement, or *lift*, is $\approx 81.5/64.1 = 1.27$.

Take-home lessons

To get predicted classes from a model, we often have to threshold predicted probabilities.

- Seems like 0.5 is a reasonable baseline.
- But other thresholds might be appropriate for some problems.

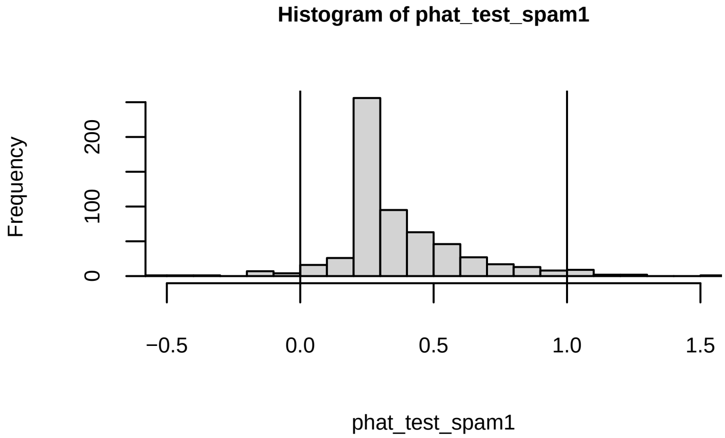
To measure the accuracy of the model, we might simply ask about its overall classification accuracy ("0/1 loss"). But we'll see other model evaluation metrics that are more nuanced.

Comparing a model to a baseline or "null model" is often an important sanity check, especially in complicated problems.

- The null model might be one that knows nothing about x.
- OR it might be a very simple model.

LPM: illegal probabilities

The linear probability model has one obvious problem: it can produce fitted probabilities that fall outside (0,1). E.g. here is a histogram of predicted probabilities for the spam test set, where 34/601 predictions (5.6%) have this problem:



LPM: illegal probabilities

This is a bit inelegant, and it happens for a straightforward reason.

Recall the basic form of the linear probability model:

$$P(y = 1 \mid x) = x \cdot \beta$$

The core of the problem is this:

- the left-hand side needs to be constrained to fall between 0 and 1, by the basic rules of probabilities
- but the right-hand side is unconstrained – it can be any real number.

Modifying the LPM

A natural fix to this problem is to break our model down into two pieces:

$$P(y = 1 \mid x) = g(x \cdot \beta)$$

The inner piece, $f(x) = x \cdot \beta$, is called the *linear predictor*. It maps features x_i onto real numbers.

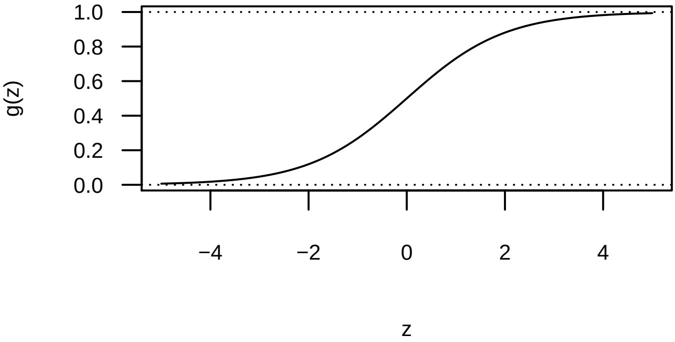
The outer piece, $g(z)$ is called a *link function.*

- It links the linear predictor $z_i \equiv f(x_i) = x_i \cdot \beta$ on the right to the probability on the left.
- It should map real numbers onto the unit interval (0,1).

Logistic regression

A standard choice is $g(z) = e^z/(1+e^z)$.

- At $z = 0$, $g(z) = 0.5$.
- When $z \to \infty$, $g(z) \to 1$, and when $z \to \infty$, $g(z) \to 0$.



Logistic regression

This is called the "logistic" or "logit" link, and it leads to the logistic regression model:

$$P(y = 1 \mid x) = \frac{\exp(x \cdot \beta)}{1 + \exp(x \cdot \beta)}$$

This is a very common choice of link function. One reason is interpretability: a little algebra shows that

$$\begin{aligned} \log \left[\frac{p}{1-p} \right] &= x \cdot \beta \\ \frac{p}{1-p} &= e^{x \cdot \beta} \end{aligned}$$

so that it is a log-linear model for the *odds* of a yes outcome.

Logistic regression is easy in R

The R syntax is nearly identical to `lm`:

```
glm(y ~ x, data=mydata, family=binomial)
```

`glm` stands for "generalized linear model," i.e. a linear model with a link function. The argument `family=binomial` tells R that y is binary and defaults to the logit link.

The response can take several forms:

- `y = 0, 1, 1,...` numeric vector
- `y = FALSE, TRUE, TRUE,...` logical
- `y = 'not spam', 'spam', 'spam',...` factor with 2 levels

Everything else is the same as in linear regression!

Logistic regression in your inbox

Let's fit a logit model to the spam data.

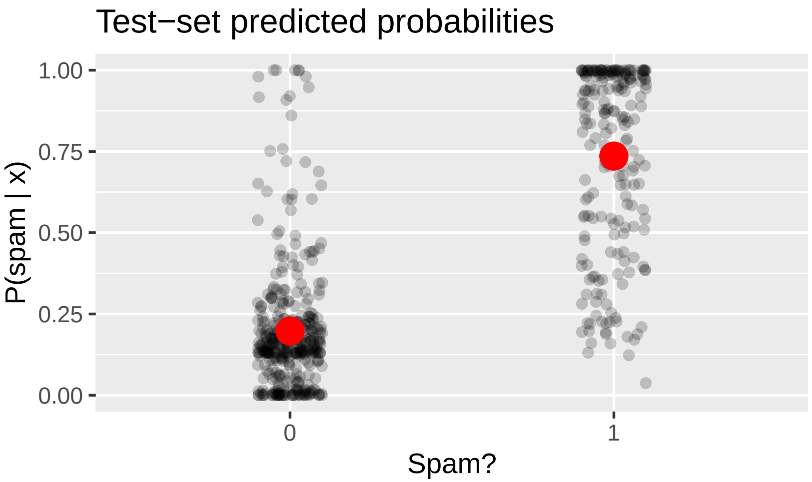
```
# Recall: the dot (.) says "use all variables not otherwise named"
logit_spam = glm(y ~ ., data=spamfit, family='binomial')
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

We're warned that some emails are clearly spam or not spam.

- This means p = 0 or p=1 up to floating-point numerical precision for at least one training-set example.
- This `warning` is largely benign and isn't something to worry about.
- However, if **all or most** of the predicted probabilities are that extreme, then either your problem is really easy or your model might be overfitting the training data.

Logistic regression in your inbox



Interpeting coefficients in LR

```r
coef(logit_spam) %>% round(2)
```

```
##                (Intercept)            word.freq.remove
##                      -2.24                        5.72
##           word.freq.order              word.freq.free
##                       0.92                        1.10
##         word.freq.meeting                word.freq.re
##                      -3.91                       -0.38
##             word.freq.edu         char.freq.semicolon
##                      -1.26                       -1.69
##     char.freq.exclamation capital.run.length.average
##                       2.23                        0.34
```

Recall our model is

$$\text{Odds} = \frac{p}{1-p} = e^{\beta_0} \cdot e^{\beta_1 x_1} \cdots e^{\beta_p x_p}$$

So e^{β_j} is an *odds multiplier* or *odds ratio* for for a one-unit increase in feature x_j.

Interpeting coefficients in LR

```r
coef(logit_spam) %>% round(2)
```

```
##                 (Intercept)            word.freq.remove
##                       -2.24                        5.72
##             word.freq.order              word.freq.free
##                        0.92                        1.10
##           word.freq.meeting                word.freq.re
##                       -3.91                       -0.38
##               word.freq.edu         char.freq.semicolon
##                       -1.26                       -1.69
##      char.freq.exclamation capital.run.length.average
##                        2.23                        0.34
```

The β for `char.freq.free` is 1.1. So having an extra `free` in an e-mail multiplies odds of spam by $e^{1.1} \approx 3$.

Interpeting coefficients in LR

```
coef(logit_spam) %>% round(2)
```

```
##                (Intercept)           word.freq.remove
##                      -2.24                       5.72
##            word.freq.order             word.freq.free
##                       0.92                       1.10
##          word.freq.meeting               word.freq.re
##                      -3.91                      -0.38
##              word.freq.edu        char.freq.semicolon
##                      -1.26                      -1.69
##       char.freq.exclamation capital.run.length.average
##                       2.23                       0.34
```

The β for `char.freq.semicolon` is -1.7. So having an extra semicolon in an e-mail multiplies odds of spam by $e^{-1.7} \approx 0.2$. (Down by a factor of five! Note to spammers: use more complex syntax.)

Interpeting coefficients in LR

```r
coef(logit_spam) %>% round(2)
```

```
##                (Intercept)            word.freq.remove
##                      -2.24                        5.72
##            word.freq.order              word.freq.free
##                       0.92                        1.10
##          word.freq.meeting                word.freq.re
##                      -3.91                       -0.38
##             word.freq.edu         char.freq.semicolon
##                      -1.26                       -1.69
##       char.freq.exclamation capital.run.length.average
##                       2.23                        0.34
```

The β for `word.freq.remove` is 5.7. So having an extra `remove` in an e-mail multiplies odds of spam by $e^{5.7} \approx 300$.

Q: What is the odds multiplier for a coefficient of 0?

LR for spam: out-of-sample

Let's go to the confusion matrix to check our out-of-sample performance.

```
phat_test_logit_spam = predict(logit_spam, spamtest, type='response')
yhat_test_logit_spam = ifelse(phat_test_logit_spam > 0.5, 1, 0)
confusion_out_logit = table(y = spamtest$y,
                            yhat = yhat_test_logit_spam)
confusion_out_logit
```

```
##    yhat
## y     0   1
##   0 358  27
##   1  51 165
```

We did better than the linear probability model:

- Error rate $(51+27)/601 \approx 13\%$, or accuracy of 87%.
- Absolute improvement over LPM: $87 - 81.5 = 6.5\%$.
- Lift over LPM: $87/81.5 \approx 1.07$.

Estimating a logit model

A logistic regression model is fit by the principle of **maximum likelihood**: *choose the parameters so that the observed data looks as likely as possible.*

Many fitting methods in machine learning are based on either maximum likelihood, or very similar principles. Let's see the details for this model.

Estimating a logit model

Recall from Prob/Stat that the likelihood function is the predicted probability of the observed data, as a function of the parameters.

In logistic regression, the likelihood is built from three assumptions/pieces:

1. The individual outcomes are binary.
2. The predicted probabilities are related to the model parameters via
$$P(y_i = 1 \mid x_i) = \frac{e^{x \cdot \beta}}{1 + e^{x \cdot \beta}}$$
$$P(y_i = 0 \mid x_i) = 1 - \frac{e^{x \cdot \beta}}{1 + e^{x \cdot \beta}} = \frac{1}{1 + e^{x \cdot \beta}}$$
3. Each binary outcome is presumed independent of the others.

Estimating a logit model

Let's think about the likelihood for a single observation (the ith one). This answers the question: how likely was it that we saw this particular outcome ($y = 0$ or $y = 1$) for observation i, assuming the true parameter was β?

Here's a convenient way to write it:

$$L_i(\beta) = \left(\frac{e^{x \cdot \beta}}{1 + e^{x \cdot \beta}}\right)^{y_i} \cdot \left(\frac{1}{1 + e^{x \cdot \beta}}\right)^{1 - y_i}$$

If $y_i = 1$, the second term gets zeroed out. Similarly, if $y_i = 0$, the first term gets zeroed out.

Estimating a logit model

Now we invoke independence. The overall likelihood is then

$$L(\beta) = \prod_{i=1}^{N} L_i(\beta)$$

or on a log scale, to avoid numerical underflow:

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{N} \log L_i(\beta) \\ &= \sum_{i=1}^{N} \left[y_i \cdot x_i \cdot \beta - \log(1 + e^{x \cdot \beta}) \right] \end{aligned}$$

This quantity can be maximized as a function of β using an iterative numerical routine (typically Newton's method, sometimes gradient ascent or BFGS). Details for another course (feel free to ask me)!

Error rates

We can take a slightly more nuanced look at the performance of a classifier than simply calculating an overall accuracy/error rate.

Here are three simple metrics you should know about:

- true positive rate (sensivity, recall)
- the false positive rate (specificity)
- the false discovery rate (precision, positive predictive value)

Let's see these three in action on our logit classifier for spam.

LR for spam: true positive rate

The *true positive rate* (TPR): among spam e-mails ($y = 1$), how many are correctly flagged as spam ($\hat{y} = 1$)?

```
##    yhat
## y     0   1
##   0 358  27
##   1  51 165
```

Here the out-of-sample TPR is $165/(51 + 165) \approx 0.76$.

Synonyms for the TPR: sensitivity, recall.

LR for spam: false positive rate

The *false positive rate* (FPR): among non-spam e-mails ($y = 0$), how many are wrongly flagged as spam ($\hat{y} = 1$)?

```
##     yhat
## y      0   1
##    0 358  27
##    1  51 165
```

Here the out-of-sample FPR is $27/(27 + 358) \approx 0.07$.

Synonyms: *specificity* is the opposite of FPR, but conveys same information:

$$\text{Specificity} = 1 - \text{FPR}$$

So this procedure had a 93% out-of-sample specificity.

LR for spam: false discovery rate

The *false discovery rate* (FDR): among e-mails flagged as spam ($\hat{y} = 1$), how many were actually not spam ($y = 0$)?

```
##      yhat
## y       0   1
##    0 358  27
##    1  51 165
```

Here the out-of-sample FDR is $27/(27 + 165) \approx 0.14$.

Synonyms: The *precision*/*positive predictive value* is the opposite of FDR, but convey same information:

$$\text{Precision} = \text{Positive Predictive Value} = 1 - \text{FDR}$$

So this procedure had a 86% precision. Among flagged spam e-mails, 86% were actually spam.

Who uses these terms?

All these synonyms for the same error rates can be a pain! But their usage tends to be field-dependent.

- FPR, FNR, FDR: statistics, machine learning
- Sensivity, specificity, positive predictive value: medicine, epidemiology, and public health
- Precision and recall: database and search engine design, machine learning, computational linguistics

Solution: always go back to the confusion matrix! It tells the whole story. Ironically, the confusion matrix *avoids confusion* over terminology.

ROC curve

In our discussion our these error rates for our spam classifier, we use a threshold of 50%.

- $P(y = 1|x) \geq 0.5 \longrightarrow$ spam
- $P(y = 1|x) < 0.5 \longrightarrow$ not spam

But what if we varied the threshold?

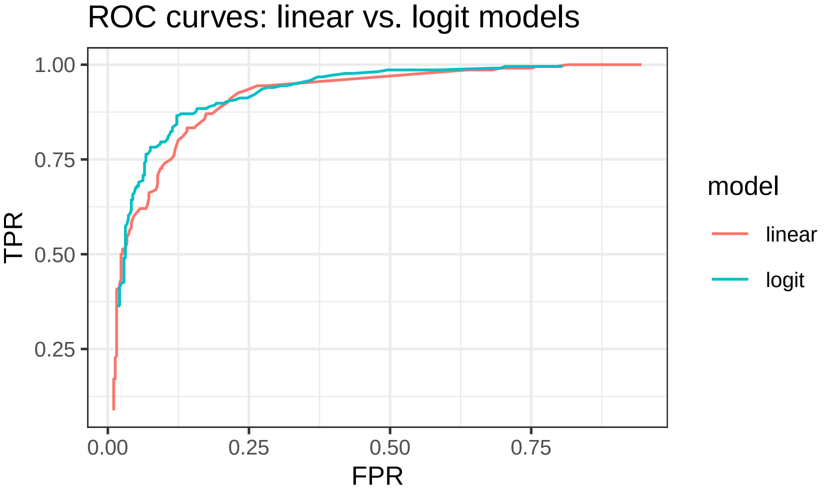
This is the question addressed by a **ROC curve:**

- a ROC ("receiver operating characteristic")[1] curve is a graph showing the performance of a binary classifier at all classification thresholds.
- at each threshold t we compute both the FPR and the TPR.
- we then graph TPR(t) versus FPR(t) as t varies.

[1] This name comes from radar operators in WWII.

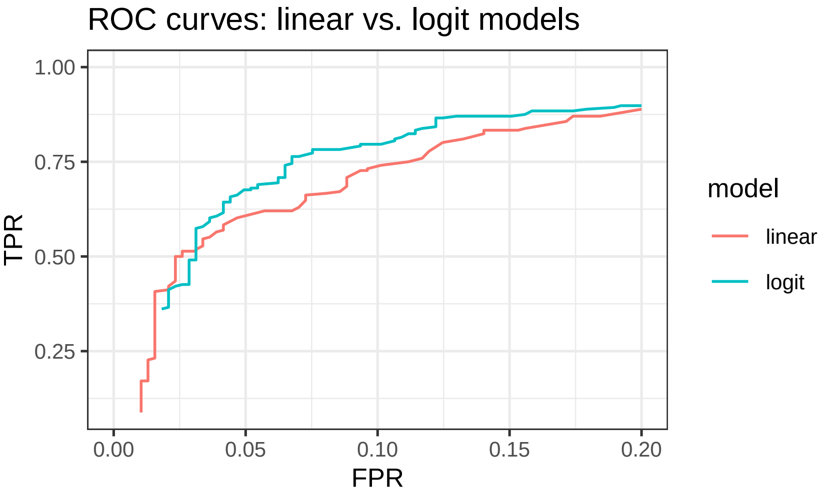
ROC curve for spam classifiers

Let's look at the ROC curve for our two spam classifiers



ROC curve for spam classifiers

Zoomed in for FPR between 0 and 0.2.



ROC curve: summary

A ROC curve plots TPR(t) vs. FPR(t) as functions of the classification threshold t.

- Thus in calculus lingo, it is a "parametric representation," where we choose to define the curve's x and y values in terms of another variable (here, the threshold) for simplicity.
- Lowering the classification threshold classifies more items as positive (y=1).
- This increases both False Positives and True Positives.

A ROC curve that is more "up and to the left" represents better performance, i.e. better detection of true positives at a fixed false positive rate.

Some people report the area under the ROC curve (AUC) as an overall measure of classifier performance.

KNN for classification

Let's see another approach to classification: K-nearest-neighbors. Super intuitive:

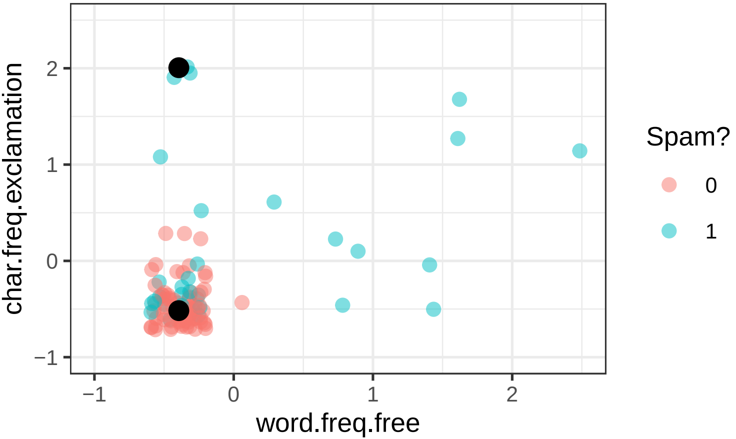
- Suppose we want to predict the class for some new $x^\star$.
- Let's ask: what is the most common class for training-set observations around $x^\star$?

We have to measure nearness using some metric, typically Euclidean distance:

$$d(x, x') = \sqrt{\sum_{j=1}^{p} (x_j - x'_j)^2}$$

Remember the importance of scaling your feature variables here! Typically we use distances scaled by $\mathrm{sd}(x_j)$ rather than raw distances.

KNN for classification



KNN for classification

In-class example: classifying glass shards for a recycling center

6 classes:

- WinF: float glass window
- WinNF: non-float window
- Veh: vehicle window
- Con: container (bottles)
- Tabl: tableware
- Head: vehicle headlamp

See `glass.R` on the class website!

Limitations of KNN for classification

Nearest-neighbor classification is simple, but limited.

- There is no good way to choose K. Train/test splits work, but they are unstable: different data $\longrightarrow$ different K (perhaps *very* different).
- The classification can be very sensitive to K.
- All you get is a classification, with only rough probabilities. E.g. with $k = 5$, all probability estimates are multiple of 20%. Without accurate probabilities, it is hard to assess misclassification risk.
- But the basic idea is the same as in logistic regression: Observations with similar x's should be classified similarly.

Multinomial logistic regression

In logistic regression, we get binary class probabilities.

In multi-class problems, the response is one of K categories. We'll encode this as $y_i = [0, 0, 1, \ldots, 0]$ where $y_{ik} = 1$ if response i is in class $k \in \{1, \ldots, K\}$.

In multinomial logistic regression (MLR), we fit a model for

$$E(y_{ik} \mid x_i) = P(y_{ik} = 1 \mid x_i) = g(x_i \cdot \beta_k)$$

That is, we fit regression coefficients for *each class.*

Multinomial logistic regression

In the MLR model, we construct this by analogy with the sigmoid link function (from binary LR) as follows:

$$\hat{p}_{ik} = P(y_{ik} = 1 \mid x_i) = \frac{e^{x_i \cdot \beta_k}}{\sum_{l=1}^{K} e^{x_i \cdot \beta_l}}$$

I like to think of this as each class vying to predict the outcome for x_i as its own, via a "rate and normalize" procedure:

- each class "rates" x_i as $e^{x_i \cdot \beta_k}$. The closer x_i is to the class-specific regression coefficient β_k, the bigger this rating is.
- Ratings $\rightarrow$ probs: divide by the sum of the ratings across classes.
- This is often called the "softmax" function.

Multinomial logit: glass example

```
library(nnet)
fgl_split = initial_split(fgl, prop=0.8)
fgl_train = training(fgl_split); fgl_test = testing(fgl_split)
ml1 = multinom(type ~ RI + Mg, data=fgl_train)
```

```
## # weights:  24 (15 variable)
## initial  value 308.182629
## iter  10 value 210.332670
## iter  20 value 190.735882
## iter  30 value 190.417408
## final  value 190.417058
## converged
```

```
coef(ml1) %>% round(2)
```

```
##       (Intercept)    RI    Mg
## WinNF        5.29 -0.16 -1.48
## Veh         -1.82 -0.13  0.18
## Con          6.72 -0.43 -3.03
## Tabl         6.23 -0.49 -2.75
## Head         7.71 -0.53 -3.24
```

Multinomial logit: glass example

Fitted class probabilities for the first five test-set examples:

```r
predict(ml1, fgl_test, type='probs') %>%
  head(5) %>%
  round(3)
```

```
##       WinF WinNF   Veh   Con  Tabl  Head
## 3  0.338 0.472 0.133 0.013 0.024 0.020
## 8  0.411 0.424 0.136 0.007 0.013 0.010
## 10 0.408 0.427 0.135 0.008 0.013 0.010
## 14 0.393 0.442 0.130 0.008 0.015 0.011
## 15 0.408 0.429 0.133 0.008 0.013 0.010
```

Multinomial logit: glass example

How did we do? Let's look at a confusion matrix:

```
yhat_test = predict(ml1, newdata = fgl_test, type='class')
conf_mat = table(fgl_test$type, yhat_test)
conf_mat
```

```
##        yhat_test
##         WinF WinNF Veh Con Tabl Head
##   WinF     7    12   0   0    0    0
##   WinNF    0    13   0   0    0    0
##   Veh      0     2   0   0    0    0
##   Con      0     2   0   0    0    1
##   Tabl     0     1   0   0    0    0
##   Head     0     2   0   0    0    2
```

```
sum(diag(conf_mat))/sum(conf_mat)
```

```
## [1] 0.5238095
```

Evaluating a classifier

In making decisions, both costs and probabilities matter. E.g. if $P(y = 1 \mid x) = 0.3$, how would you respond differently if:

- x is word content of an e-mail and y is spam status?
- x is mammogram result and y is breast cancer status?
- x is DNA test and y is guilty/not guilty?

Different kinds of errors may have different costs. Thus it helps to de-couple two tasks: *modeling probabilities accurately* and *making decisions.*

This suggests that we evaluate the performance of a classifier in terms its *predicted probabilities*, not its *decisions about class labels.*

Evaluating a classifier: likelihood

The natural way to do us is by calculating the *likelihood* for our model's predicted probabilities. Suppose that our classifier produces predicted probabilities $\hat{p}_{ik}$ for each response i and class k. Then the likelihood is

$$\begin{aligned} \text{Like} &= \prod_{i=1}^{n} \prod_{l=1}^{K} \hat{p}_{il}^{y_{il}} \\ &= \prod_{i=1}^{n} \hat{p}_{i,k_i} \end{aligned}$$

where k_i is the observed class label for case i.

To get from the first to the second lines, notice that $y_{il} = 1$ for $l = k_i$, and zero otherwise.

Evaluating a classifier: log likelihood

On a log scale, this becomes

$$\text{loglike} = \sum_{i=1}^{n} \log \hat{p}_{i,k_i}$$

In words: we sum up our model's predicted log probabilities for the outcomes y_{i,k_i} that actually happened.

As with everything in statistical learning: we can calculate an in-sample or a out-of-sample log likelihood, and the out-of-sample is more important!

Q: what's the largest possible log likelihood for a classifier?

Evaluating a classifier: deviance

Sometimes we quote a model's *deviance* instead of its log likelihood. The relationship is simple:

$$\text{deviance} = -2 \cdot \text{loglike}$$

Log likelihood measures *fit* (which we want to maximize), deviance measures *misfit* (which we want to minimize).

So the negative sign makes sense. But why the factor of 2? *Because of the analogy because least squares and the normal distribution.*

Evaluating a classifier: deviance

Remember back to an ordinary regression problem with normally distributed errors, $y_i \sim N(f(x_i), \sigma^2)$:

$$\text{Like} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}(y_i - f(x_i))^2\right\}$$

On a log scale, up to a constant not involving $f(x)$, this becomes:

$$\text{loglike} \propto -\frac{1}{2}\sum_{i=1}^{n}(y_i - f(x_i))^2 = -\text{RSS}/2$$

where RSS = residual sums of squares.

Deviance generalizes the notion of "residual sums of squares" to non-Gaussian models.

Bayes' Rule for classification

Recall Bayes' rule:

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

You might remember that each of these terms has a name:

- $P(A)$: the prior probability
- $P(A \mid B)$: the posterior probability
- $P(B \mid A)$: the likelihood
- $P(B)$: the marginal (total/overall) probability

In classification, "A" is a class label and "B" is a set of features.

Bayes' Rule for classification

Bayes's rule:

$$P(y = k \mid x) = \frac{P(y = k) \cdot P(x \mid y = k)}{P(x)}$$

$P(y = k)$ is the prior probability for class k. We usually get this from the raw class frequencies in the training data. For example:

```
table(fgl_train$type) %>% prop.table %>% round(3)
```

```
##
##  WinF WinNF   Veh   Con  Tabl  Head
## 0.297 0.366 0.087 0.058 0.047 0.145
```

Bayes' Rule for classification

Bayes's rule:

$$P(y = k \mid x) = \frac{P(y = k) \cdot P(x \mid y = k)}{P(x)}$$

$P(x)$ is the marginal probability of observing feature vector x. Notice it doesn't depend on k! It's the same number for all classes.

Thus we usually write the posterior probabilities up to this constant of proportionality, without bothering to compute it:

$$P(y = k \mid x) \propto P(y = k) \cdot P(x \mid y = k)$$

(Note: often we do the actual computations on a log scale instead.)

Bayes' Rule for classification

Bayes's rule:

$$P(y = k \mid x) = \frac{P(y = k) \cdot P(x \mid y = k)}{P(x)}$$

The hard part is estimating the likelihood $P(x \mid y = k)$. In words: how likely is it that we would have observed feature vector x if the true class label were k?

This is like regression in reverse!

Naive Bayes

Recall that $x = (x_1, x_2, \ldots, x_p)$ is a vector of p features. The simplest strategy for estimating $P(x \mid y = k)$ is called "Naive Bayes."

It's "naive" because we make the simplifying assumption that *every feature x_j is independent* of all other features, conditional on the class labels:

$$\begin{aligned} P(x \mid y = k) &= P(x_1, x_2, \ldots, x_p \mid y = k) \\ &= \prod_{j=1}^{p} P(x_j \mid y = k) \quad \text{(independence)} \end{aligned}$$

This simplifies the requirements of the problem: *just calculate the marginal distribution of the features,* i.e. $P(x_j \mid y = k)$ for all features j and classes k.

Naive Bayes: a small example

In `congress109.csv` we have data on all speeches given on the floor of the U.S. Congress during the 109th Congressional Session (January 3, 2005 to January 3, 2007).

Every row is a set of *phrase counts* associated with a single representative's speeches across the whole session. X_{ij} = number of times that rep i utter phrase j during a speech.

The target variable $y \in$ R, D is the party affiliation of the representative.

Naive Bayes: a small example

We'll focus on just a few phrases and famous politicians:

```
# read in data
congress109 = read.csv("../data/congress109.csv", header=TRUE, row.names=1)
congress109members = read.csv("../data/congress109members.csv", header=TRUE, row.names=1)
```

Focus on a few key phrases and a few famous pols:

```
X_small = dplyr::select(congress109, minimum.wage, war.terror, tax.relief, hurricane.katrina)
X_small[c('John McCain', 'Mike Pence', 'John Kerry', 'Edward Kennedy'),]
```

```
##                minimum.wage war.terror tax.relief hurricane.katrina
## John McCain               0         27          0                14
## Mike Pence                0         12          1                11
## John Kerry               12         16         13                23
## Edward Kennedy          260          8          1                53
```

Naive Bayes: a small example

Let's look at these counts summed across all members in each party:

```
y = congress109members$party

# Sum phrase counts by party
R_rows = which(y == 'R')
D_rows = which(y == 'D')
colSums(X_small[R_rows,])
```

```
##      minimum.wage        war.terror        tax.relief hurricane.katrina
##               294               604               497               717
```

```
colSums(X_small[D_rows,])
```

```
##      minimum.wage        war.terror        tax.relief hurricane.katrina
##               767               237               176              1295
```

So we get the sense that some phrases are "more Republican" and some "more Democrat."

Naive Bayes: a small example

To make this precise, let's build our Naive Bayes model for a Congressional speech:

- Imagine that every phrase uttered in a speech is a random sample from a "bag of phrases," where each phrase has its own probability. (*This is the Naive Bayes assumption of independence.*)
- Here the bag consists of just four phrases: "minimum wage", "war on terror", "tax relief," and "hurricane katrina".
- Each class (R or D) has its own probability vector associated with the phrases in the bag.

Naive Bayes: a small example

We can estimate these probability vectors for each class from the phrase counts in the training data.

For Republicans:

```
probhat_R = colSums(X_small[R_rows,])
probhat_R = probhat_R/sum(probhat_R)
probhat_R %>% round(3)
```

```
##      minimum.wage        war.terror        tax.relief hurricane.katrina
##             0.139             0.286             0.235             0.339
```

And for Democrats:

```
probhat_D = colSums(X_small[D_rows,])
probhat_D = probhat_D/sum(probhat_D)
probhat_D %>% round(3)
```

```
##      minimum.wage        war.terror        tax.relief hurricane.katrina
##             0.310             0.096             0.071             0.523
```

Naive Bayes: a small example

Let's now look at some particular member of Congress and try to build the likelihood, $P(x \mid y)$, for his or her phrase counts

```
X_small['Sheila Jackson-Lee',]
```

```
##                    minimum.wage war.terror tax.relief hurricane.katrina
## Sheila Jackson-Lee           11         15          3                66
```

Are Sheila Jackon-Lee's phrase counts $x = (11, 15, 3, 66)$ more likely under the Republican or Democrat probability vector?

Naive Bayes: a small example

Recall $P(x \mid y = R)$:

```
##     minimum.wage        war.terror        tax.relief hurricane.katrina
##           0.1392            0.2860            0.2353            0.3395
```

Under this probability vector:

$$
\begin{aligned}
P(x \mid y = \mathrm{R}) &= P(x_1 = 11 \mid y = \mathrm{R}) \\
&\times P(x_2 = 15 \mid y = \mathrm{R}) \\
&\times P(x_3 = 3 \mid y = \mathrm{R}) \\
&\times P(x_4 = 66 \mid y = \mathrm{R}) \\
&= (0.1392)^{11} \cdot (0.2860)^{15} \cdot (0.2353)^{3} \cdot (0.3395)^{66} \\
&= 3.765 \times 10^{-51}
\end{aligned}
$$

Naive Bayes: a small example

Now recall $P(x \mid y = D)$:

```
##      minimum.wage        war.terror        tax.relief hurricane.katrina
##            0.1392            0.2860            0.2353            0.3395
```

Under this probability vector:

$$
\begin{aligned}
P(x \mid y = \mathrm{D}) &= P(x_1 = 11 \mid y = \mathrm{D}) \\
&\times P(x_2 = 15 \mid y = \mathrm{D}) \\
&\times P(x_3 = 3 \mid y = \mathrm{D}) \\
&\times P(x_4 = 66 \mid y = \mathrm{D}) \\
&= (0.3099)^{11} \cdot (0.0958)^{15} \cdot (0.0711)^{3} \cdot (0.5232)^{66} \\
&= 1.293 \times 10^{-43}
\end{aligned}
$$

Naive Bayes: a small example

These numbers are tiny, so it's much safer to work on a log scale:

$$\log P(x \mid y = k) = \sum_{j=1}^{p} x_j \log p_j^{(k)}$$

where $p_j^{(k)}$ is the jth entry in the probability vector for class k.

```
x_try = X_small['Sheila Jackson-Lee',]
sum(x_try * log(probhat_R))
```

```
## [1] -116.1083
```

```
sum(x_try * log(probhat_D))
```

```
## [1] -98.75633
```

Naive Bayes: a small example

Let's use Bayes' rule (posterior $\propto$ prior times likelihood) to put this together with our prior, estimated using the empirical class frequencies:

```
table(y) %>% prop.table %>% round(3)
```

```
## y
##     D     I     R
## 0.457 0.004 0.539
```

So:

$$P(R \mid x) \propto 0.539 \cdot (3.765 \times 10^{-51})$$

and

$$P(D \mid x) \propto 0.457 \cdot (1.293 \times 10^{-43})$$

Naive Bayes: a small example

To actually calculate a posterior, we must turn this into a set of probabilities by normalizing, i.e. dividing by the sum across all classes:

$$
\begin{aligned}
P(D \mid x) &= \frac{0.457 \cdot (1.293 \times 10^{-43})}{0.457 \cdot (1.293 \times 10^{-43} + 0.539 \cdot (3.765 \times 10^{-51})} \\
&\approx 1
\end{aligned}
$$

So:

1. Our model thinks Sheila Jackson-Lee is a Democrat.
2. The data completely overwhelm the prior! This is often the case in Naive Bayes models.

Naive Bayes: a bigger example

Let's turn to `congress109_bayes.R` to see a larger example of Naive Bayes classification, where we fit our model with all 1000 phrase counts.

Naive Bayes: summary

- Works by directly modeling $P(x \mid y)$, versus $P(y \mid x)$.
- This **regression in reverse** only works because we assume that each feature in x is independent, given the class labels.
- Simple and easy to compute, and therefore scalable to very large data sets and classification problems.
- Unlike a logit model, it works even more with features P than examples N.
- Often too simple: the "naive" assumption of independence really is a drastic simplification.
- The resulting probabilities are useful for classification purposes, but often not believeable as probabilities.
- Most useful when the features x are categorical variables (like phrase counts!) Very common in text analysis.