Class Notes

Statistical Computing & Machine Learning

Class 10

Classification overview

Response variable: categorical. Typically just a few levels: 2 or 3.

Two types of outputs from models:

1. The predicted category given the inputs
2. Probability of each category given the inputs

   Type (2) can be fitted with maximum likelihood.

Trade-offs:

- Flexibility vs interpretability
- Accuracy vs bias

Four model architectures

1. Logistic regression. Especially important for interpretability.
2. Linear discriminant analysis
3. Quadratic discriminant analysis
4. K nearest neighbors

Today

1. Probability and odds
   - Theme Song
   - Making book
2. Multivariate gaussians

Probability and odds

Probability $p(event)$ is a number between zero and one.

Simple way to make a probability model for yes/no variable: encode outcome as zero and one, use regression.

```
Whickham$alive <- as.numeric(with(Whickham, outcome == "Alive"))
```

Model of mortality in Whickham

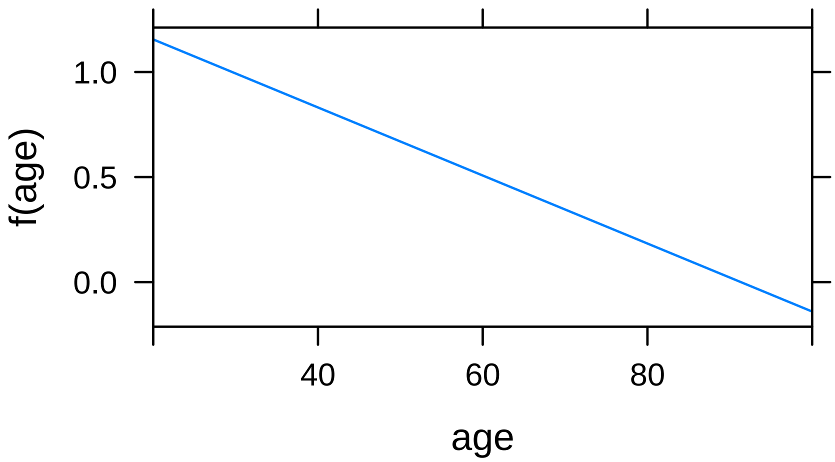
```
res <- mean( alive ~ smoker, data=Whickham)
res
```

```
##        No       Yes
## 0.6857923 0.7611684
```

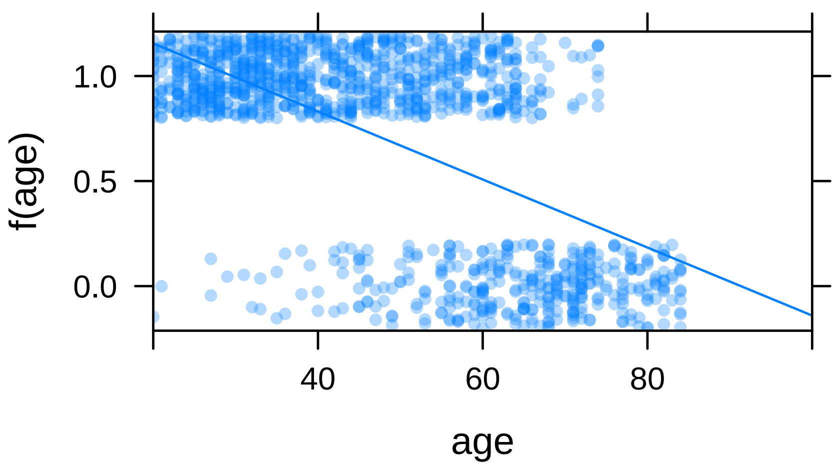
```
res / (1-res)
```

```
##       No      Yes
## 2.182609 3.187050
```

```
mod2 <- lm(alive ~ age, data=Whickham)
f <- makeFun(mod2)
plotFun(f(age) ~ age, age.lim = c(20,100))
```




```
plotPoints(jitter(alive) ~ age, data=Whickham, add=TRUE,
           pch=20, alpha=.3)
```



If we're going to use likelihood to fit, the estimated probability can't be ≤ 0.

Log Odds

Gerolamo Cardano (1501-1576) defined *odds* as the ratio of favorable to unfavorable outcomes.

For an event whose *probability* is p, it's *odds* are $w = \frac{p}{1-p}$.

A probability is a number between 0 and one.

An odds is a ratio of two positive numbers. 5:9, 9:5, etc.

"Odds are against it," could be taken to mean that the odds is less than 1. More unfavorable outcomes than favorable ones.

Given odds w, the probability is $p = \frac{w}{1+w}$. There's a one-to-one correspondence between probability and odds.

The log odds is a number between $-\infty$ and ∞.

Why use odds?

Making Book

Several horses in a race. People bet on each one amounts H_i.

What should be the winnings when horse j wins? Payoff means you get your original stake back plus your winnings.

If it's arranged to pay winnings of

$\sum i \neq j \frac{H_i}{H_j}$ + the amount H_j

the net income will be zero for the bookie.

Shaving the odds means to pay less than the zero-net-income winnings.

Link function

You can build a linear regression to predict the log odds, $\ln w$. The output of the linear regression is free to range from $-\infty$ to ∞. Then, to measure likelihood, unlog to get odds w, then $p = \frac{w}{1+w}$.

Use of glm()

Response should be 0 or 1. We don't take the log odds of the response. Instead, the likelihood is

- p if the outcome is 1 - $1-p$ if the outcome is 0

Multiply these together of all the cases to get the total likelihood.

Interpretation of coefficients

Each adds to the log odds in the normal, linear regression way. Negative means less likely; positive more likely.

Multivariate Gaussian

Joint probabilities and classification

Suppose we have K classes, $A_1, A_2, \ldots, A_K$. We also have a set of inputs $x_1, x_2, \ldots, x_p := \mathbf{x}$.

We observe $\mathbf{x}$ and we want to know $p(A_j|\mathbf{x})$.

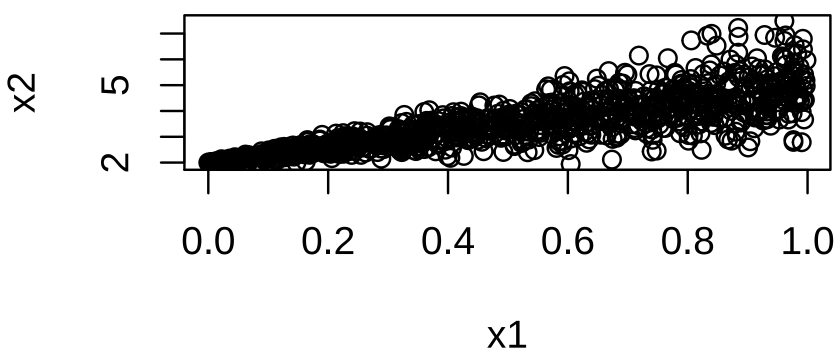
To set things up so that we can find $p(A_j|\mathbf{x})$, we collect a lot of objects of class A_j and measure $\mathbf{x}$ from each of them. We use this to create a model probability:

$$p(\mathbf{x}|A_j)$$

Independent variables x_i

Describing dependence

```
x1 = runif(1000)
x2 = rnorm(1000, mean=3*x1+2, sd=x1)
plot(x1, x2)
```



Linear correlations and the Gaussian

Remember the univariate Gaussian with parameters μ and σ^2:

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

In-class programming activity

Fitting a logistic regression link