Lecture 6: Introduction to Classification

Will Welch (adapted from materials by Matias Salibian-Barrera)

Department of Statistics, University of British Columbia

STAT 447B: Methods for Statistical Learning

September–December 2014

Today's Learning Goals

1. Examples of Classification

2. Review of Logistic Regression

Vaso Constriction Data

```
> help(vaso, package='robustbase')

(...)

 A data frame with 39 observations on the following 3 variables.
 'Volume' Inhaled volume of air
 'Rate' Rate of inhalation
 'Y' vector of 0 or 1 values.

Details:

The data taken from Finney (1947) were obtained in a carefully
controlled study in human physiology where a reflex 'vaso
constriction' may occur in the skin of the digits after taking a
single deep breath.  The response y is the occurrence (y = 1) or
non-occurrence (y = 0) of vaso constriction in the skin of the
digits of a subject after he or she inhaled a certain volume of
air at a certain rate.  The responses of three subjects are
available.  The first contributed 9 responses, the second
contributed 8 responses, and the third contributed 22 responses.
```

Spam e-mail data

From the UCI Machine Learning Repository

https://archive.ics.uci.edu/ml/datasets/Spambase

GOAL: Determine whether a given email is spam or not.

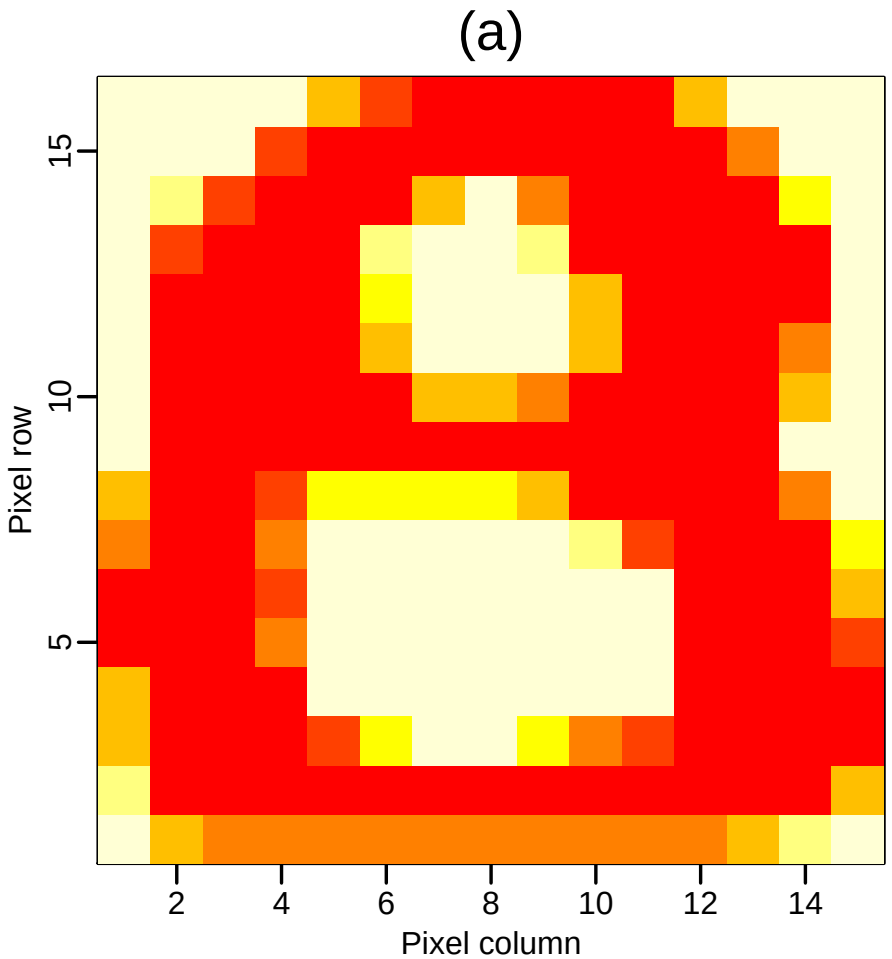
False positives (marking good mail as spam) are very undesirable.
If we insist on zero false positives in the training/testing set,
20-25% of the spam passed through the filter.

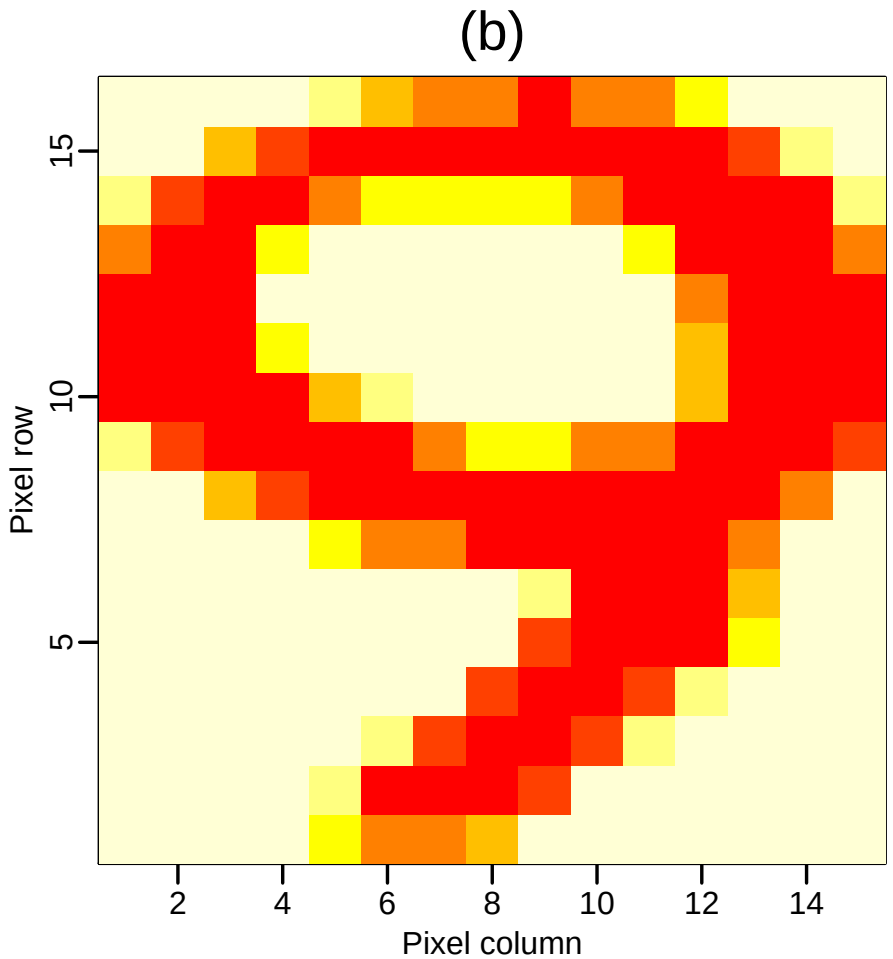
Number of Instances: 4601 (1813 Spam = 39.4%)

Number of Attributes: 57 continuous

Most of the attributes indicate whether a particular word or character was frequently occuring in the e-mail. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters.

Digit Recognition





Digit Recognition

- Again from the UCI Machine Learning Repository
  `https://archive.ics.uci.edu/ml/machine-learning-databases/mfeat/mfeat-pix`
- 10 classes, one for each of the digits $0, \ldots, 9$
- Can turn this into a 2-class problem by considering only two digits, e.g., “8” and ”9”
- 240 explanatory variables from 15×16 averages of pixels from a grey-scale image of a handwritten digit, taking values 0–7
- Database has 200 cases for each of the 10 digits (“0” data first, then “1” data, etc.)

Vaso Constriction: Logistic Regression

- Variables: Y (constriction, 0 or 1), v (Volume) and Rate
- A linear model (ignoring Rate for now):

$$E[Y|v] = \beta_0 + \beta_1 v$$

- A generalized linear model:

$$E[Y|v] = \Pr(Y = 1|v) = \frac{\exp(\beta_0 + \beta_1 v)}{1 + \exp(\beta_0 + \beta_1 v)}$$

- Likelihood when $Y_i \sim \text{Bernoulli}(1, p_i)$, independent

$$L(y_1, \ldots, y_n; p_1, \ldots, p_n) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \quad \text{where } p_i = \Pr(Y_i = 1)$$

Vaso Constriction: Logistic Regression

- The p_i are determined by only 2 parameters, β_0 and β_1

$$p_i = \Pr(Y_i = 1 | v = v_i) = \frac{\exp(\beta_0 + \beta_1 v_i)}{1 + \exp(\beta_0 + \beta_1 v_i)}$$

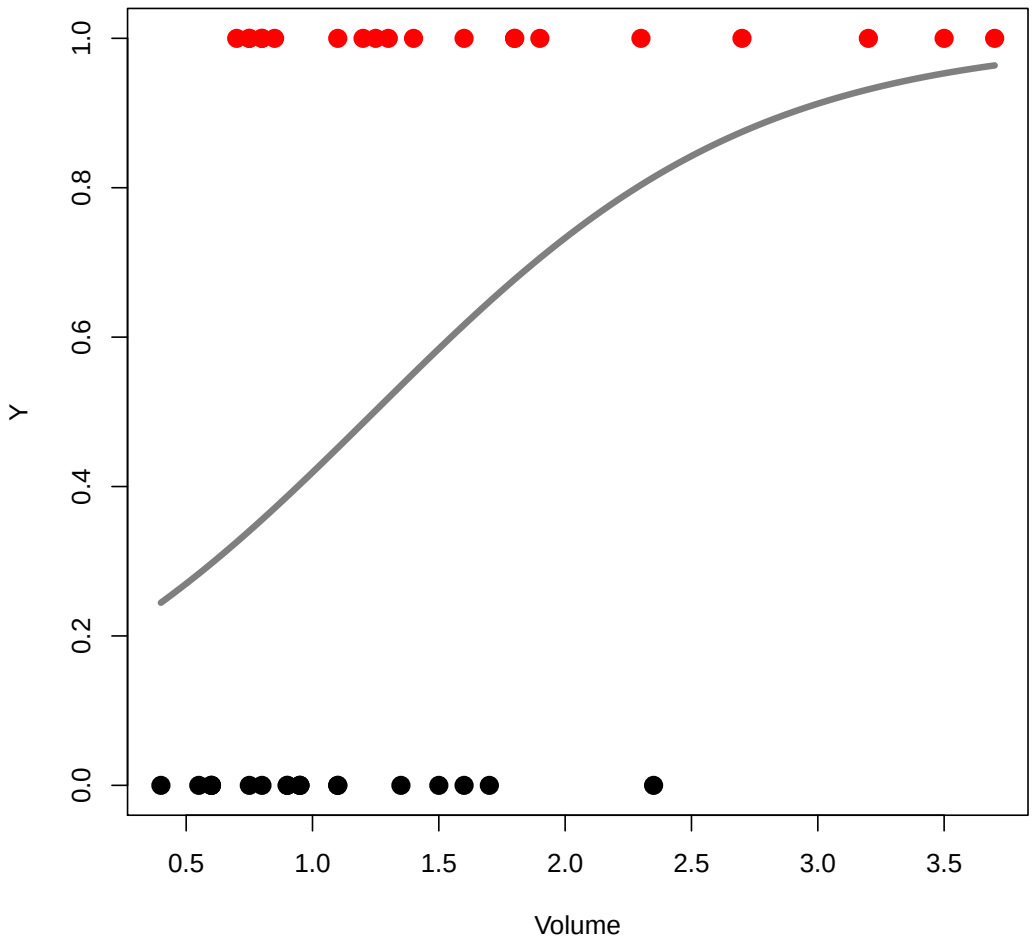
- So, we need to maximize over β_0 and β_1

$$L(y_1, \ldots, y_n; v_1, \ldots, v_n; \beta_0, \beta_1)$$

- Given a value of Volume, the predicted probability of $Y = 1$ is

$$\frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 v\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 v\right)}$$

Vaso Constriction: Logistic Regression



The curve is the prediction of $\Pr(Y = 1)$. Note the logistic shape.

Digits: Logistic Regression

- $y = 0/1$ codes "8" or "9"
- $$p_i = \frac{\exp\left(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_{240} x_{240,i}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_{240} x_{240,i}\right)}$$
- Predictions: all 400 digits classified correctly!
- Is this a valid assessment of classification accuracy?
- How long did it take?
- Stay tuned!