Logistic regression

Hastie, Tibshirani, Friedman Ch 4.4

Kevin Murphy Ch. 8

CS 6140
Machine Learning
Professor Olga Vitek

February 6, 2017

Generative vs discriminative models

- Goal: predict Y
  - Bayes rule:

$$p(Y|\mathbf{X}) = \frac{p(Y) \cdot p(\mathbf{X}|Y)}{p(\mathbf{X})}$$

- Generative classifiers
  - Specify prior probability of $p(Y)$
  - Assume conditional distribution $p(\mathbf{X}|Y)$
  - Use Bayes rule to derive the posterior $p(Y|\mathbf{X})$
  - **Example:** Linear discriminant analysis

- Discriminative classifiers
  - Estimate the posterior the posterior $p(Y|\mathbf{X})$
  - Do not assume the distribution on $\mathbf{X}$
  - **Example:** Y binary: logistic regression

Probability Distribution of a Binary Outcome Y

- In many situations, the response variable has only two possible outcomes
  - Disease ($Y = 1$) vs Not diseased ($Y = 0$)
  - Employed ($Y = 1$) vs Unemployed ($Y = 0$)
- Response is *binary or dichotomous*
- Can model response using Bernoulli dist

Y_i	Probability
1	$\Pr\{Y_1 = 1\} = \pi_i$
0	$\Pr\{Y_1 = 0\} = 1 - \pi_i$

- $E\{Y_i\} = \pi_i$
- $Var\{Y_i\} = \pi_i(1 - \pi_i)$

Goal: express $E\{Y\}$ as function of a covariate X

- The simple regression is not appropriate

$$E\{Y_i\} \;=\; \beta_0 + \beta_1 X_i$$

It violates several assumptions:

(1) Does not enforce the constraint $0 \leq E\{Y_i\} \leq 1$ is

(2) Non-normal (binary) distribution of $\varepsilon \mid X$:

When $Y_i = 0$: $\varepsilon_i = 0 - \beta_0 - \beta_1 X_i$
When $Y_i = 1$: $\varepsilon_i = 1 - \beta_0 - \beta_1 X_i$

(3) Non-constant variance

$$\begin{aligned} Var\{Y_i\} &= \pi_i(1-\pi_i) \\ &= (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i) \end{aligned}$$

Solution: a Generalized Linear Model

- A generalized linear model is

$$\begin{aligned} E\{Y_i\} &= g(\beta_0 + \beta_1 X_i), \text{ or} \\ g^{-1}(E\{Y_i\}) &= \beta_0 + \beta_1 X_i \end{aligned}$$

  where g is a sigmoid function in (0,1).

  – g is called the *mean response function*

  – g^{-1} is called the *link function*

- A choice of g produces different models

  – $g(t) = \text{Identity}$
  $\rightarrow$ linear regression

  – $g(t) = \Phi(t) =$ standard Normal CDF
  $\rightarrow$ probit regresison

  – $g(t) = \frac{\exp(t)}{1+exp(t)} =$ CDF of the logistic distrib.
  $\rightarrow$ logistic regresison

Motivation for Probit Regression: Latent Variable

- Assume that the binary response is guided by a non-observed continuous variable

- Example: linear model for blood pressure:

$$\text{bp} = \beta_0 + \beta_1 \text{age} + \varepsilon$$

Only observe

$$Y = \begin{cases} 1 \text{ (disease)}, & \text{if blood pressure} > c \\ 0 \text{ (healthy)}, & \text{if blood pressure} \leq c \end{cases}$$

$$\begin{aligned} \Pr\{Y = 1\} & \\ = & \quad \Pr\{\text{bp} > c\} = \Pr\{\beta_0 + \beta_1 \text{age} + \varepsilon > c\} \\ = & \quad \Pr\{\varepsilon < \beta_0 + \beta_1 \text{age} - c\} \\ = & \quad \Pr\{\frac{\varepsilon}{\sigma} < \frac{\beta_0 - c}{\sigma} + \frac{\beta_1}{\sigma}\text{age}\} \\ = & \quad \Pr\{z < \beta_0' + \beta_1' \text{age}\} \\ = & \quad \Phi(\beta_0' + \beta_1' \text{age}) \end{aligned}$$

Logistic Response Function and Logistic Regression

- A sigmoidal response function

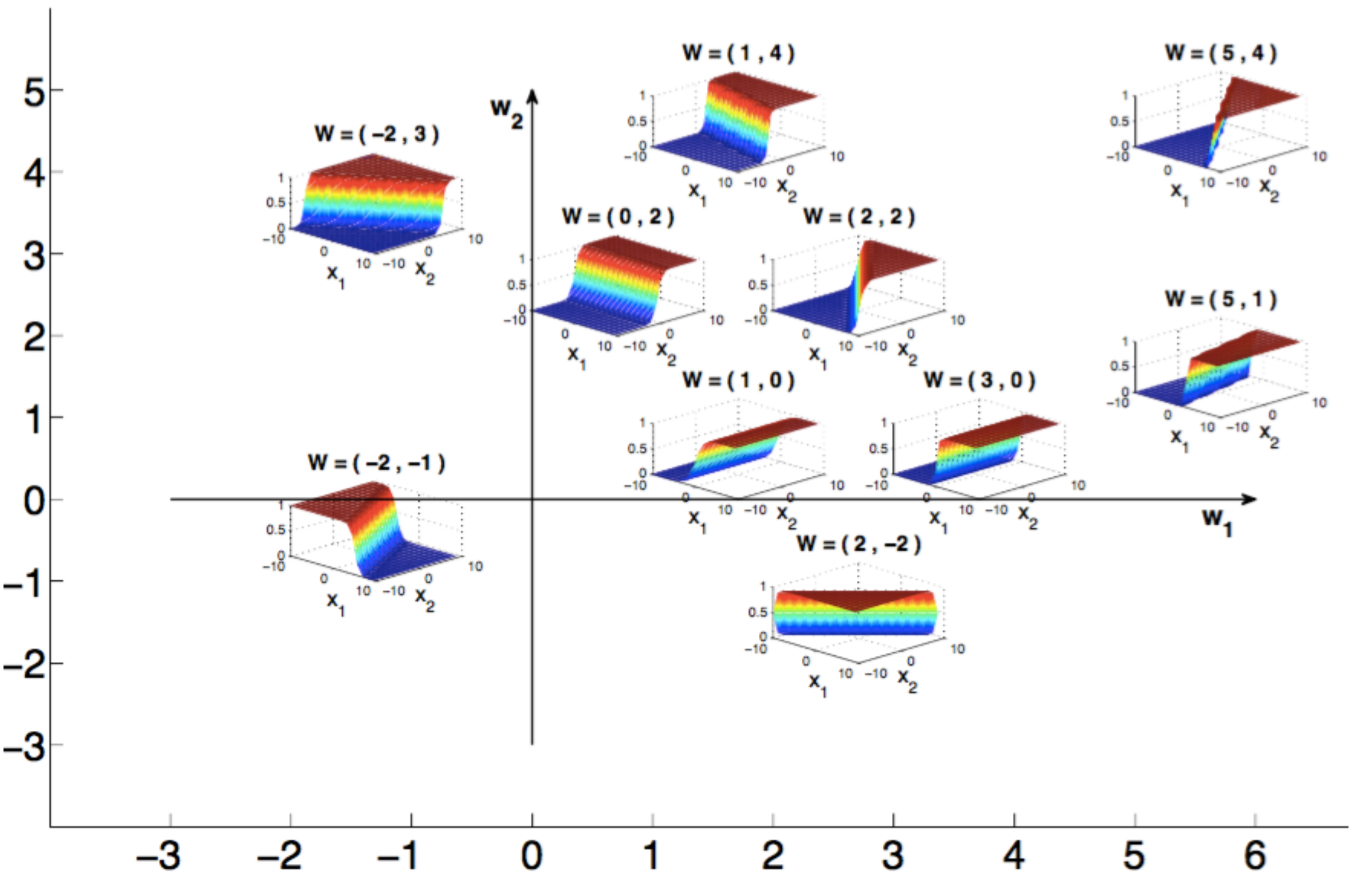
$$
\begin{aligned}
E\{Y_i\} &= \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \\
&= \frac{1}{1 + \exp(-\beta_0 - \beta_1 X_i))}
\end{aligned}
$$

  - A monotonic increasing/decreasing function
  - Explicit functional form
  - Restricts $0 \leq \mathrm{E}(Y_i) \leq 1$
  - Example of a **nonlinear** model

- **Logit** link function

$$
\log\left(\frac{E\{Y_i\}}{1 - E\{Y_i\}}\right) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 X_i
$$

Sigm($\beta_1 x_1 + \beta_2 x_2$)



K. Murphy, Fig 8.1

Probability Distribution of Y in Logistic Regression

- Y_i are independent but not identically distributed Bernoulli random variables

$$\begin{aligned} Y_i &\overset{ind}{\sim} \text{Bernoulli}(\pi_i) \text{ where} \\ \pi_i &= \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \end{aligned}$$

  - note no more error term!

- Probability density of Y_i

$$f(Y_i) = \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}$$

- Least Squares Estimates are inappropriate

  - use maximum likelihood for parameter estimation

Estimation by Maximum Likelihood

- $Y_i \overset{ind}{\sim}$ Bernoulli(π_i) where

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

- Log likelihood: $\log_e(L) =$

$$\begin{aligned}
&= \log \left\{ \prod_{i=1}^{n} \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i} \right\} \\
&= \sum_{i=1}^{n} Y_i \log(\pi_i) + \sum_{i=1}^{n} (1 - Y_i) \log(1 - \pi_i) \\
&= \sum_{i=1}^{n} Y_i \log(\frac{\pi_i}{1 - \pi_i}) + \sum_{i=1}^{n} \log(1 - \pi_i) \\
&= \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \log(1 + \exp(\beta_0 + \beta_1 X_i))
\end{aligned}$$

- MLEs do not have closed forms

Equivalent specification: Binomial distribution

- Change in notation
  - Data: (Y_{ij}, n_i, X_i), $i = 1, 2, \cdots, c$
  - X_i : predictor for observation i
  - n_i : # of Bernoulli trials in observation i
  - $Y_i' := \sum_{j=1}^{n_i} Y_{ij}$
  - Model:

$$\begin{aligned} Y_i' &\overset{ind}{\sim} Binomial(n_i, \pi_i), \text{ where} \\ \pi_i &= \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \end{aligned}$$

- Log-Likelihood: $\log_e(L) =$

$$\begin{aligned} &= \log \prod_{i=1}^{c} \left\{ \left(\begin{array}{c} n_i \\ Y_i' \end{array} \right) \pi^{Y_i'} (1 - \pi_i)^{n_i - Y_i'} \right\} \\ &= \sum_{i=1}^{c} \left\{ Y_i' \log(\pi_i) + (n_i - Y_i') \log(1 - \pi_i) + \log \left(\begin{array}{c} n_i \\ Y_i' \end{array} \right) \right\} \\ &= \sum_{i=1}^{c} \left\{ Y_i' \log \frac{\pi_i}{1 - \pi_i} + n_i \log(1 - \pi_i) + \log \left(\begin{array}{c} n_i \\ Y_i' \end{array} \right) \right\} \end{aligned}$$

Equivalence of Bernouilli and Binomial Models

- Binomial Log-Likelihood equals Bernouilli Log-Likelihood, up to a constant:

$$\log_e(L)^{Binomial} =$$

$$= \sum_{i=1}^{c} \left\{ Y_i' \log(\pi_i) + (n_i - Y_i') \log(1 - \pi_i) \right\} + constant$$

$$= \sum_{i=1}^{c} \left\{ \sum_{j=1}^{n_i} Y_{ij} \log(\pi_i) + (n_i - \sum_{j=1}^{n_i} Y_{ij}) \log(1 - \pi_i) \right\} + constant$$

$$= \sum_{i=1}^{c} \sum_{j=1}^{n_i} \left\{ Y_{ij} \log(\frac{\pi_i}{1 - \pi_i}) + \log(1 - \pi_i) \right\} + constant$$

$$= \log_e(L)^{Bernouilli} + constant$$

- Both models lead to same parameter estimates and inferences, but have different deviances.

Summaries of prediction/classification

		True class	
		p	**n**
Hypothesized class	**Y**	**T**rue **P**ositives	**F**alse **P**ositives
	N	**F**alse **N**egatives	**T**rue **N**egatives
Column totals:		**P**	**N**

$\text{fp rate} = \frac{FP}{N}$ $\text{tp rate} = \frac{TP}{P}$

$\text{precision} = \frac{TP}{TP+FP}$ $\text{recall} = \frac{TP}{P}$

$\text{accuracy} = \frac{TP+TN}{P+N}$

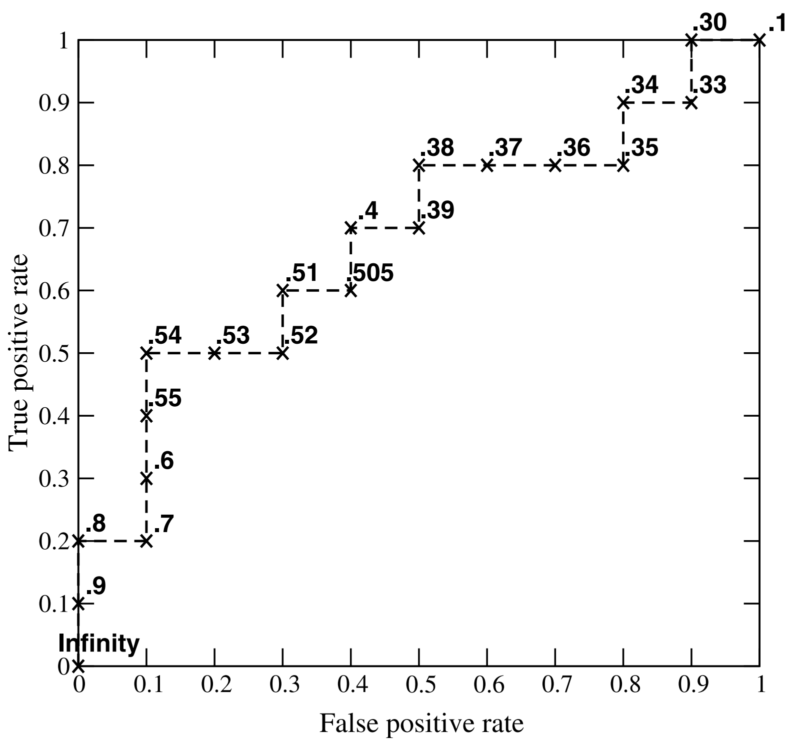
$\text{F-measure} = \frac{2}{1/\text{precision}+1/\text{recall}}$

- Results over multiple score cutoffs are summarized in a Receiver Operating Characteristic (ROC) curve
- Vary c, and for all c plot sensitivity vs 1-specificity. Evaluate models by area under the curve.
- Area = 1 $\rightarrow$ perfect classification

  Area = .5 $\rightarrow$ random classification.

Fawcett, "An introduction to ROC analysis". *Pattern Recognition Letters*, 2005

Inst#	Class	Score	Inst#	Class	Score
1	**p**	.9	11	**p**	.4
2	**p**	.8	12	**n**	.39
3	**n**	.7	13	**p**	.38
4	**p**	.6	14	**n**	.37
5	**p**	.55	15	**n**	.36
6	**p**	.54	16	**n**	.35
7	**n**	.53	17	**p**	.34
8	**n**	.52	18	**n**	.33
9	**p**	.51	19	**p**	.30
10	**n**	.505	20	**n**	.1



Fawcett, "An introduction to ROC analysis". *Pattern Recognition Letters*, 2005

Automatic Variable Selection

- Exhaustive search. Minimize:

$$\begin{aligned} & -2\log_e L(\mathbf{b}) \\ AIC_p &= -2\log_e L(\mathbf{b}) + 2p \\ BIC_p &= -2\log_e L(\mathbf{b}) + p\log_e(n) \end{aligned}$$

- Heuristic search
  - forward selection; backward elimination; stepwise selection

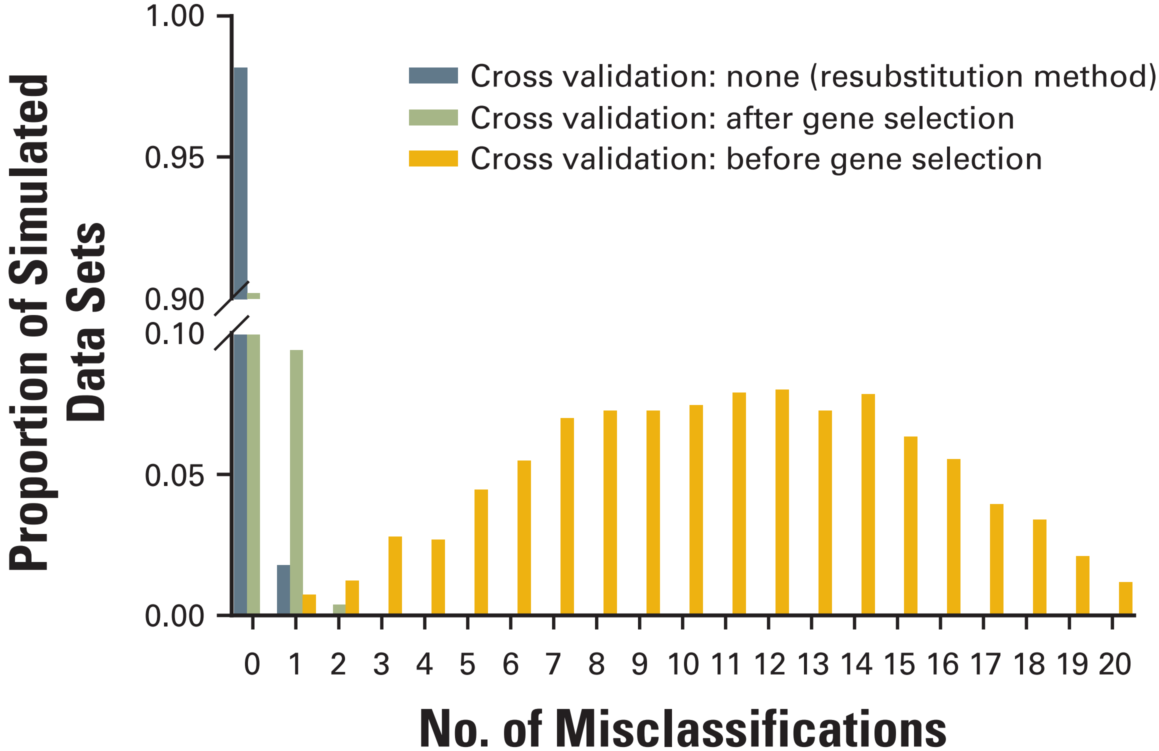
- Statistical regularization
  - Negative log-likelihood penalized with ridge, lasso, elastic net etc

Variable Selection Should be Done as Part of Cross-Validation

- Example from Simon *et al.*, JNCI, 2003.
- Simulated data with no structure
  - 20 observations with random labels
  - 6,000 possible but unrelated predictors
  - Repeated 200 times
- Estimated predictive accuracy using
  - no cross-validation
  - selecting features on full dataset, then using cross-validation
  - selecting features at each step of cross-validation

Variable Selection Should be Done as Part of Cross-Validation

Example from Simon *et al.*, JNCI, 2003.



- Conclusion
  - Incorporating selection of predictors within the cross-validation procedure is key

More than 2 groups: conditional multinomial distributions

- $Y|X = x \sim Multinomial(n_{i+}, \pi_1(x), \ldots, \pi_J(x))$
- X - predictor;
  Y - multinomial response with J categories

Row	Column 1	$\cdots$	Column J	Total
1	π_{11} $(\pi_{1\|1})$	$\cdots$	π_{1J} $(\pi_{J\|1})$	π_{1+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	π_{I1} $(\pi_{1\|I})$	$\cdots$	π_{IJ} $(\pi_{J\|I})$	π_{I+}
Total	π_{+1}	$\cdots$	π_{+J}	π_{++}

- Consider the new notation:
  $\pi_j(X) = P(Y = j|X = x)$, $\sum_{j=1}^{J} \pi_j = 1$
- If Y is ordered, we can also be interested in cumulative probabilities:
  $P_j(X) = P(Y \leq j|X = x)$

Baseline-Category Logistic Regression

- *Data:*
  (y_{ij}, x_i); $y_{ij} = I_{\{y_i=j\}}$; $\sum_{j=1}^{J} y_{ij} = 1$;
  $j = 1, \ldots, J,\ i = 1, \ldots, n$

- The model describes the effects of covariates X on the $J-1$ logits.

$$\log \frac{\pi_j(X)}{\pi_1(X)} = \alpha_j + \beta_j X,\ j = 2, \ldots, J$$

  - An arbitrary category ($j = 1$ or $j = J$) is chosen as the baseline category
  - Each other category j is paired with the baseline to build a logistic model
  - Separate set of parameters β_j for each π_j
  - Separate linear relationship between X and $log\frac{\pi_j(X)}{\pi_1(X)}$
  - Values of β_j depend on the baseline

Predicted Probability

- Since $\pi_j(X) = \pi_1(X)e^{\alpha_j + \beta_j X}$ and $\sum_{j=1}^{J} \pi_j(X) = 1$

$$\begin{cases} \pi_1(X) = \dfrac{1}{1+\sum_{k=2}^{J} \exp(\alpha_k + \beta_k X)}, \ j = 1 \\ \\ \pi_j(X) = \dfrac{\exp(\alpha_j + \beta_j X)}{1+\sum_{k=2}^{J} \exp(\alpha_k + \beta_k X)}, \ j = 2, 3, \ldots, J \end{cases}$$

- Same denominator for all j

- For $J = 2$ - an ordinary logistic regression

- Can be viewed as a classification model
  - the observation is assigned to the category with the highest predicted probability
  - can plot ROC curves for a particular category versus all other categories combined

Max. Likelihood Estimation

Since

$$\pi_1(x_i) = [1 + \sum_{j=2}^{J} exp(\alpha_j + \beta_j x_i)]^{-1} \; and \; y_{i1} = (1 - \sum_{j=2}^{J} y_{ij}),$$

contribution of (y_{ij}, x_i) to the log-likelihood is:

$$\begin{aligned} l_i = log \left[\prod_{j=1}^{J} \pi_j(x_i)^{y_{ij}} \right] & \\ = & \sum_{j=2}^{J} y_{ij} \; log \pi_j(x_i) + \left(1 - \sum_{j=2}^{J} y_{ij} \right) \; log \pi_1(x_i) \\ = & \sum_{j=2}^{J} y_{ij} \; log \frac{\pi_j(x_i)}{1 - \sum_{k=2}^{J} \pi_k(x_i)} + log \pi_1(x_i) \\ = & \sum_{j=2}^{J} y_{ij} \; (\alpha_i + \beta_j x_i) - log \left[1 + \sum_{j=2}^{J} exp(\alpha_i + \beta_j x_i) \right] \end{aligned}$$

- Maximize $\sum_{i=1}^{I} l_i$ with respect to α_j and β_j

Multinomial Vs Binary Logistic Regression

- $\log \frac{\pi_j(X)}{\pi_1(X)} = \alpha_j + \beta_j X,\ j = 2, \ldots, J$

- Can we fit separate $J-1$ logistic regressions for $J-1$ response categories vs baseline?
  - Same model in principle

- Separate-fitting ML parameter estimates
  - Differ from the joint-fitting ML estimates
  - Tend to have larger standard errors
  - Loss of efficiency is minor when the baseline is the most common category

Example: Dose Response

```
#---------------read the data----------------------
x <- data.frame(
    count = c(59, 25, 46, 48, 32, 48, 21, 44, 47, 30, 44,
              14, 54, 64, 31, 43, 4, 49, 58, 41),
    dose = c(rep(0, 5), rep(1, 5), rep(2, 5), rep(3, 5)),
    response = as.factor(rep(c(0:4), 4))
)
m <- matrix(x$count, byrow=TRUE, ncol=5,
     dimnames=(list(0:3, 0:4)))
> m
   0  1  2  3  4
0 59 25 46 48 32
1 48 21 44 47 30
2 44 14 54 64 31
3 43  4 49 58 41
```

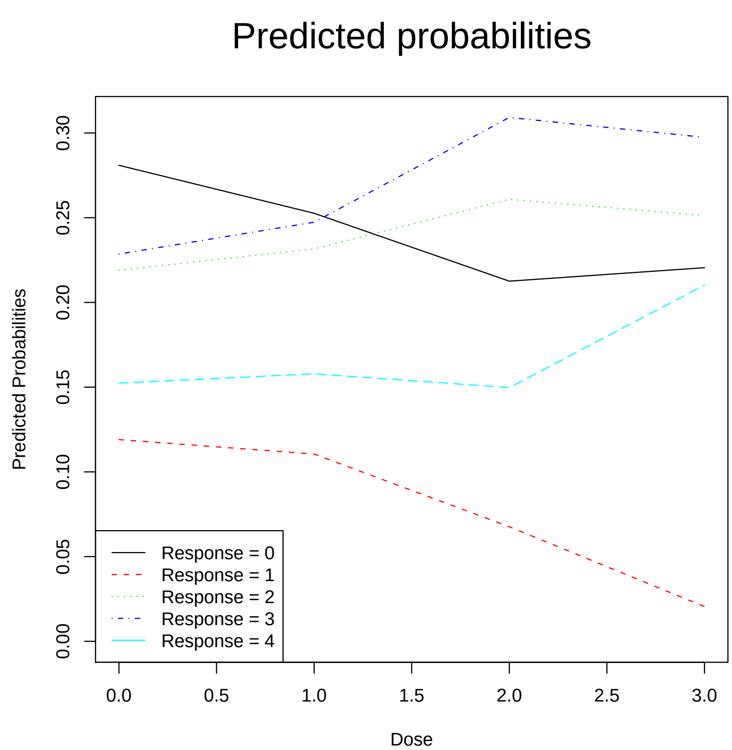
- Row='dose'; column='response'
- View 'dose' as a categorical predictor
  - introduce 3 indicators for predictors
- View 'dose' as a continuous predictor
  - assign scores to categories on an arbitrary scale
- More R examples: Venable & Ripley p. 203

Example: Dose Response 'Dose' Viewed as Categorical

```
library(nnet)
fit1 <- multinom(response ~ as.factor(dose), weights=count,
   data=x)

> summary(fit1)
Coefficients:
  (Intercept)  ...(dose)1  ...(dose)2  ...(dose)3
1  -0.8586335  0.03194971  -0.2864809  -1.5161958
2  -0.2488754  0.16185828   0.4536705   0.3794879
3  -0.2063195  0.18526707   0.5810140   0.5055581
4  -0.6117850  0.14178037   0.2615807   0.5641507

Std. Errors:
  (Intercept)  ...(dose)1  ...(dose)2  ...(dose)3
1   0.2386396   0.3541205   0.3887204   0.5746170
2   0.1966936   0.2867909   0.2827264   0.2869711
3   0.1943777   0.2826526   0.2759257   0.2797853
4   0.2195434   0.3199468   0.3212239   0.3095891

Residual Deviance: 2443.166
AIC: 2475.166
```

Plot Predicted Probabilities

```
predProb <- unique(fit1$fitted.values)
> predProb
        0          1         2         3         4
 0.2809484 0.11904927 0.2190490 0.2285720 0.1523813
 0.2526320 0.11052594 0.2315780 0.2473691 0.1578949
 0.2125601 0.06763396 0.2608694 0.3091786 0.1497580
 0.2205135 0.02051445 0.2512809 0.2974355 0.2102557

matplot(predProb)
legend("bottomleft", lty=c(1:4), col=c(1:5),
    paste("Response =", c(0:4)))
```



Example: Dose Response 'Dose' Viewed as Continuous

```
library(nnet)
fit2 <- multinom(response ~ dose, weights=count, data=x)
> summary(fit2)
Coefficients:
  (Intercept)       dose
1  -0.6999134 -0.3544346
2  -0.2194566  0.1470232
3  -0.1772963  0.1945578
4  -0.6544057  0.1914772
...
Residual Deviance: 2449.145
AIC: 2465.145
```