Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Supervised Learning: Linear Method (2/2)

Applied Multivariate Statistics – Spring 2012

Overview

- Logistic Regression
- Bayes rule for general loss functions

Generalized Linear Models

- Stochastic part

$$X \sim F(\theta)$$

- Deterministic part

$$\underbrace{g(\theta)}_{\text{Link function}} = \underbrace{\eta(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p}_{\text{Linear predictor}}$$

Examples

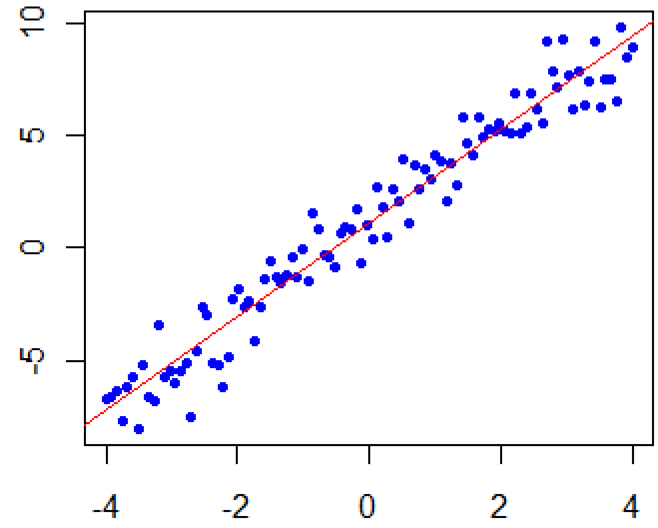
- Linear Regression

  $$Y \sim N(\mu, \sigma^2)$$

  $$\mu = \beta_0 + \beta_1 x_1$$

  Link function: Identity function

  Example: Distance and Travel time in tram

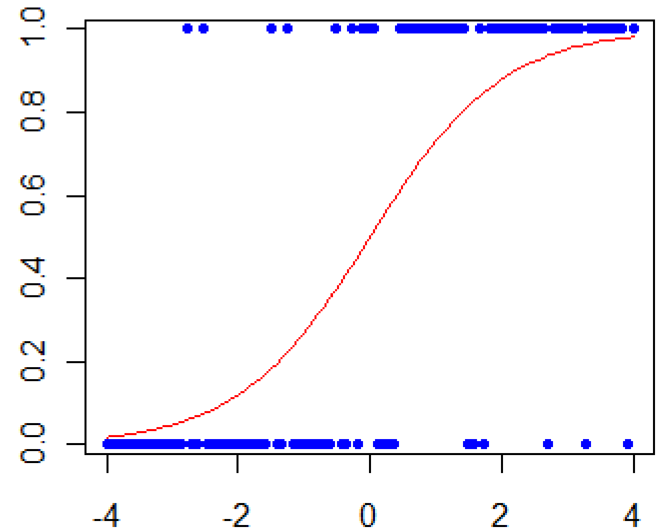


- Logistic Regression

  $$Y \sim Bernoulli(p)$$

  $$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1$$

  Link function: logit
  Example: Survival and dose of poison



Logistic regression for supervised learning

- Logistic regression computes posterior probability of class membership
- Can be used in the same way as LDA

Logistic regression and LDA are almost the same thing

- LDA: Assuming same normal density in each group

$$
\begin{aligned}
& \log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) = \\
= \quad & \underbrace{\log\left(\frac{\pi_0}{\pi_1}\right) - \frac{1}{2}(\mu_0+\mu_1)^T\Sigma^{-1}(\mu_1-\mu_0)}_{\alpha_0} + x^T\underbrace{\Sigma^{-1}(\mu_1-\mu_0)}_{\alpha} = \\
= \quad & \alpha_0 + \alpha^T x
\end{aligned}
$$

- Logistic regression by assumption:

$$
\log\left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)}\right) = \beta_0 + \beta^T x
$$

Difference between LDA and Logistic Regression

- Parameter estimate LDA:
  Maximize joint likelihood

$$\prod_i f(x_i, y_i) = \underbrace{\prod_i f(x_i|y_i)}_{\text{Gaussian}} \underbrace{\prod_i f(y_i)}_{\text{Bernoulli}}$$

- Parameter estimate Logistic Regression:
  Maximize conditional likelihood

$$\prod_i f(x_i, y_i) = \underbrace{\prod_i f(y_i|x_i)}_{\text{logistic}} \underbrace{\prod_i f(x_i)}_{\text{ignored}}$$

- Logistic Regression is thus based on less assumptions, i.e., more flexible

LDA vs. Logistic Regression

+ very comfortable implementation (CV, LD’s)

+ easy to apply to several groups

- needs more assumptions

- less comfortable implementation (CV harder, no LD’s)

- Possible but harder to use for several groups

+ needs less assumptions

Personal suggestion:

- LDA for several groups, low-dim representation, quick solutions
- Logistic Regression for two groups, applications where performance is crucial

Example: Spam Filter

- R: Function “glm” with option “family = binomial”

Loss functions

True class

Estimated class

- Loss function: L(k,l)
- Common choice: 0-1 loss

	T = 0	T = 1	T = 2
E = 0	0	1	1
E = 1	1	0	1
E = 2	1	1	0

- Other choices possible

	T = 0	T = 1	T = 2
E = 0	0	10	3
E = 1	9	0	27
E = 2	4	5	0

Mathematical background

- Classifier $c(X): X \rightarrow \{1, \ldots, k\}$
- C: true class
- Probability of miss-classification:

$$pmc(k) = P(c(X) \neq k | C = k)$$

- ***Risk function*** R for classifier c:

$$R(c, k) = E_X\left[L\big(k, c(X)\big)\big|C = k\right] =$$

$$= \sum_{l=1}^{K} L(k, l) P(c(X) = l | C = k)$$

Assuming 0-1-loss: $R(c, k) = pmc(k)$

- ***Total risk*** for classifier c:

$$R(c) = E_C[R(c, C)] = \sum_{k=1}^{K} \pi_k R(c, k)$$

Overall missclassification error

Assuming 0-1-loss: $R(c) = \sum_{k=1}^{K} \pi_k \, pmc(k)$

Bayes rule for classification

- Classification rule that minimizes total risk under 0-1-loss is

$$c(X) = argmax_{l \leq k} P(C = l | X = x)$$

- Classification rule that minimizes total risk under general loss function is

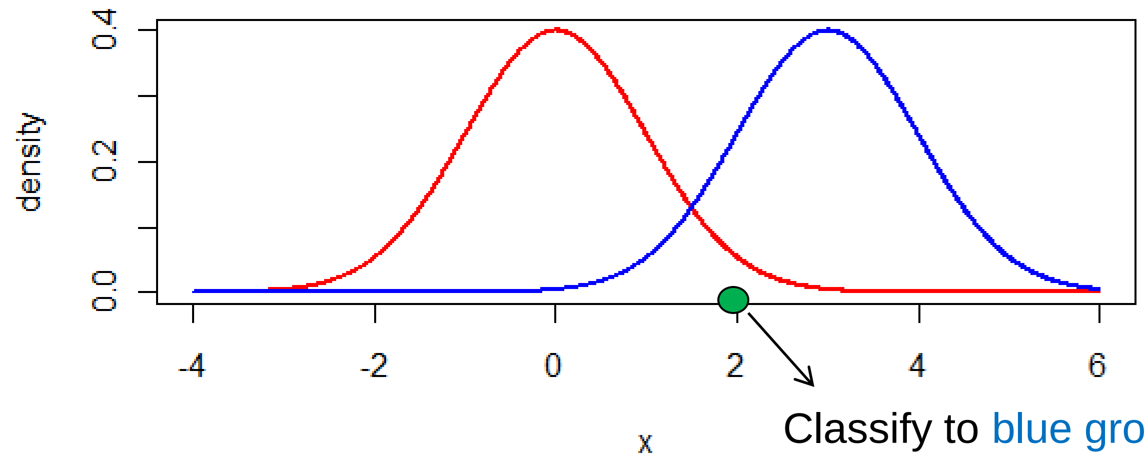
$$c(X) = argmin_{l \leq K} \sum_{j=1}^{K} L(j, l) P(C = j | X = x)$$

Bayes rule is a benchmark

- No method can beat the Bayes rule, even given an infinite amount of data; i.e., sometimes, perfect classification is not possible

  Intuition:

  Assume equal prior: Classify to group with larger density



Classify to blue group,
might still be red group

- Our job in practice: Find best possible estimate for posterior probability

Example: Detecting HIV
Assuming 0-1-loss

- Suppose LDA or Logistic regression yield for a patient $P(HIV = 0|X=x) = 0.9$, thus $P(HIV = 1|X=x) = 0.1$
- Assuming 0-1-loss

	T=HIV	T=No HIV
E=HIV	0	1
E=No HIV	1	0

- Bayes rule: Choose class HIV=0 if $P(HIV=0|X=x) > 0.5$

- Thus in example, choose HIV=0, i.e. “patient has no HIV” Total risk based on 0-1-loss will be optimal

Example: Detecting HIV
Assuming more realistic loss function

- Suppose LDA or Logistic regression yield for a patient
  P(HIV = 0|X=x) = 0.9, thus P(HIV = 1|X=x) = 0.1
- Assuming

	T=HIV	T=No HIV
E=HIV	0	1
E=No HIV	100	0

- Bayes rule: Choose class HIV=0 if

$$\sum_{j=0}^{1} L(j,0)P(HIV = j\,|X = x) < \sum_{j=0}^{1} L(j,1)P(HIV = j\,|X = x)$$

	T=HIV	T=No HIV
E=HIV	0	1
E=No HIV	100	0

Example: Continued

truth estimate

$$L(0,0)P(0|x) + L(1,0)P(1|x) < L(0,1)P(0|x) + L(1,1)P(1|x)$$
$$0 \ * P(0|x) + 100 * P(1|x) < 1 * P(0|x) + 0 * P(1|x)$$
$$100 * P(1|x) < P(0|x)$$

- Using $P(1|x) = 1 \ - P(0|x)$ we get:

$$100 \ - 100 * P(0|x) < P(0|x)$$
$$P(0|x) > \frac{100}{101} = 0.99$$

- Bayes rule: Choose class HIV=0 if P(HIV=0|X=x) > 0.99
I.e., only declare "no HIV" if you are really, really sure!

- Thus in example choose HIV=1, i.e., "patient has HIV"
Total risk based on given loss function is optimized

Concepts to know

- Logistic regression
- LDA vs. Logistic regression
- Bayes rule
  - as a benchmark
  - as a optimal rule for general loss functions

R functions to know

- Function “glm” with option family = “binomial”