Linear classification methods

Lecture 2a

Overview

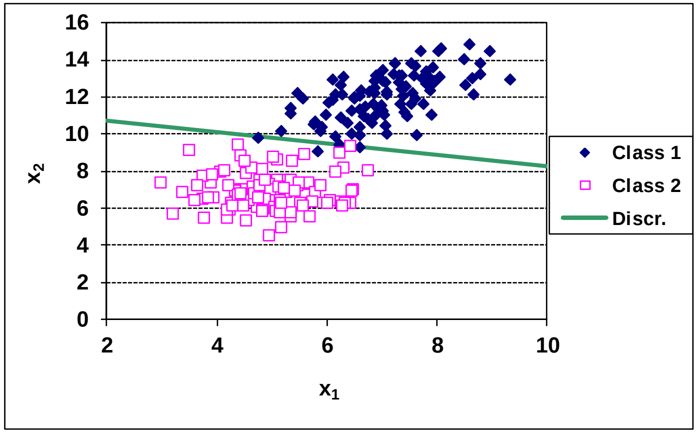
- Discriminant Analysis models
- Logistic regression
- Generalized Linear Model

Classification

- Given data $D = \left((X_i, Y_i), i = 1 \dots N\right)$
  - $Y_i = Y(X_i) = C_j \in \boldsymbol{C}$
  - Class set $\boldsymbol{C} = (C_1, \dots, C_K)$

Classification problem:

- Decide $\hat{Y}(x)$ that maps **any** x into some class C_K
  - Decision boundary



Classifiers

- **Deterministic**: decide a rule that directly maps X into $\hat{Y}$

- **Probabilistic:** define a model for $P(Y = C_i|X), i = 1 \ldots K$

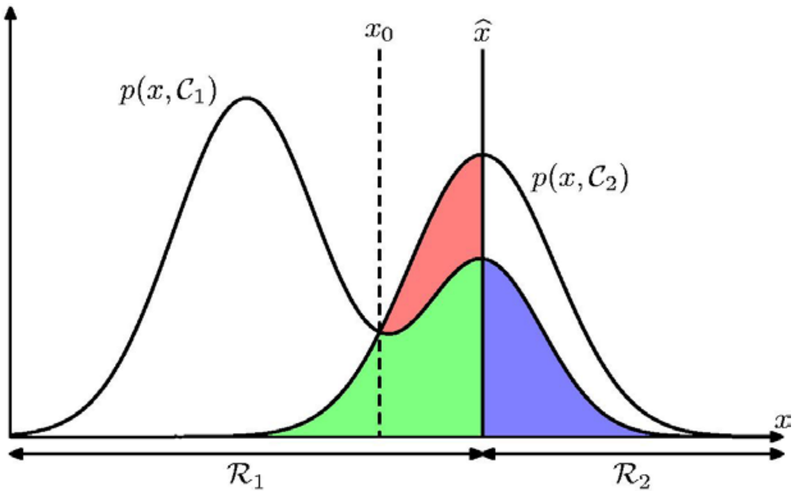
Disanvantages of deterministic classifiers:

- Sometimes simple mapping is not enough (risk of cancer)
- Difficult to embed loss-> rerun of optimizer is often needed
- Combining several classifiers into one is more problematic
  - Algorithm A classifies as spam, Algorithm B classifies as not spam → ???
  - P(Spam|A)=0.99, P(Spam|B)=0.45 → better decision can be made

Probabilities into decision

- Loss minimization

$$\min_{\hat{f}} EL(y, \hat{f}) = \min_{\hat{f}} \int L(y, \hat{f}) p(y, x|w) dxdy$$



When loss is

$$\begin{cases} 1, wrongly\ classified \\ 0, correctly\ classified \end{cases}$$

Classify Y as

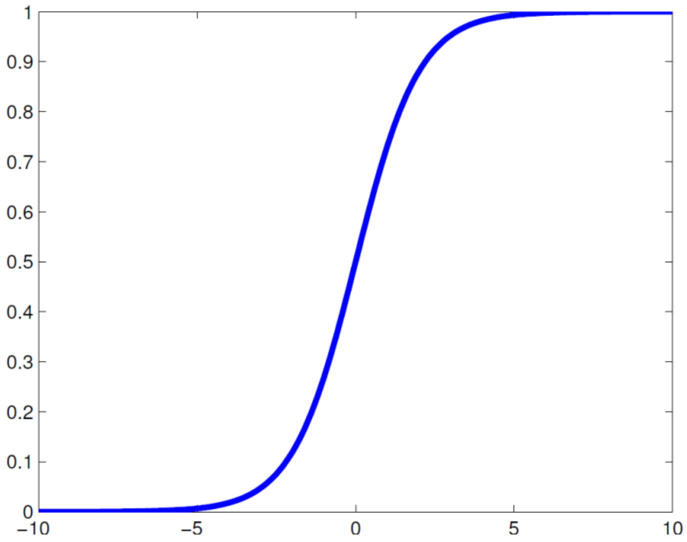
$$\hat{Y} = \arg\max_c p(Y = c|X)$$

Logistic regression

- Discriminative model
- Model for binary output
  - $C = \{C_1 = 1, C_2 = 0\}$
    $p(Y = C_1|X) = sigm(\boldsymbol{w}^T \boldsymbol{x})$

$$sigm(a) = \frac{1}{1 + e^{-a}}$$

What is $P(Y = C_2|X)$?



- Alternatively

$$Y \sim Bernoulli(sigm(a)), a = \boldsymbol{w^T x}$$

$$sigm(a) = \frac{1}{1 + e^{-a}}$$

Logistic regression

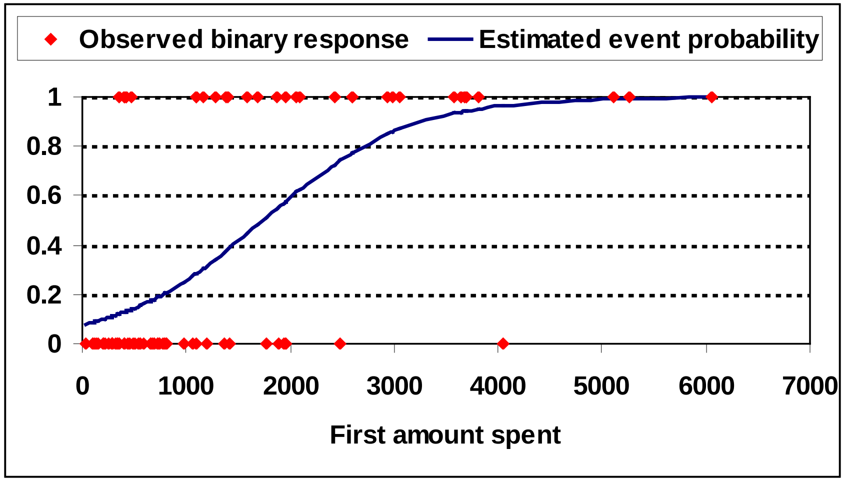
- Logistic model- yet another form

$$\ln\frac{p(Y=1|X=x)}{P(Y=0|X=x)} = \ln\frac{p(Y=1|X=x)}{1-P(Y=1|X=x)} = logit(p(Y=1|X=x)) = \boldsymbol{w}^T\boldsymbol{x}$$

The log of the odds is linear in *x*

- Here $logit(t) = \ln\left(\frac{t}{1-t}\right)$
- Note $p(Y|X)$ is connected to $\boldsymbol{w}^T\boldsymbol{x}$ via logit link

Example: Probability to buy more than once as function of First Amount Spend



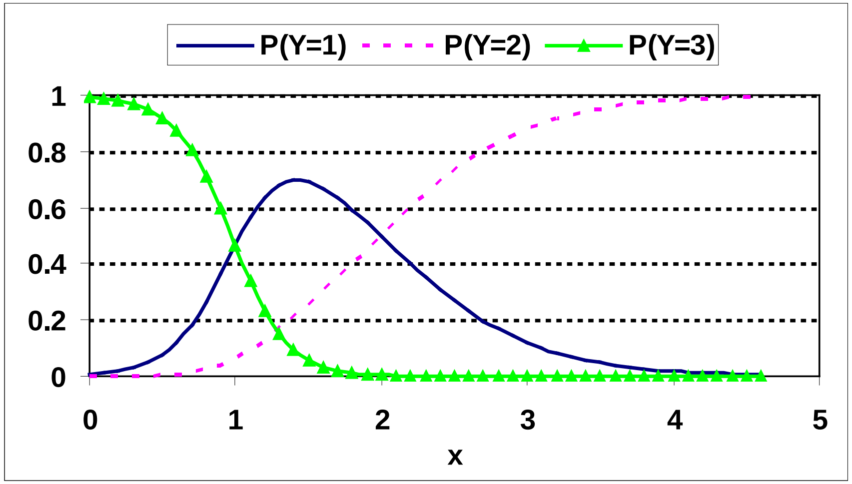
Logistic regression

- When Y is categorical,

$$p(Y = C_i|x) = \frac{e^{\boldsymbol{w}_i^T \boldsymbol{x}}}{\sum_{j=1}^{K} e^{\boldsymbol{w}_j^T \boldsymbol{x}}} = softmax(\boldsymbol{w}_i^T \boldsymbol{x})$$

- Alternatively

$$Y \sim Multinoulli\left(softmax(\boldsymbol{w}_1^T \boldsymbol{x}), \ldots softmax(\boldsymbol{w}_K^T \boldsymbol{x})\right)$$



Logistic regression

Fitting logistic regression

- In binary case,

$$\log P(D|w) = \sum_{i=1}^{N} y_i \log(\,sigm(w^T x_i)) + (1 - y_i) \log(1 - sigm(w^T x_i))$$

  - Can not be maximized analytically, but unique maximizer exists
- To maximize loglikelihood, optimization used
  - Newton's method traditionally used (Iterative Reweighted Least Squares)
  - Steepest descent,Quasi-newton methods…

Estimation:

For new x, estimate $p(y) = [p_1, \ldots p_C]$ and classify as arg $\max_i p_i$

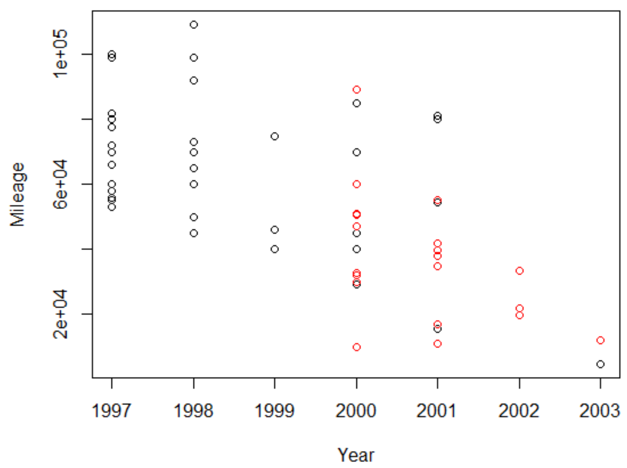
Decision boundaries of logistic regression are linear

Logistic regression

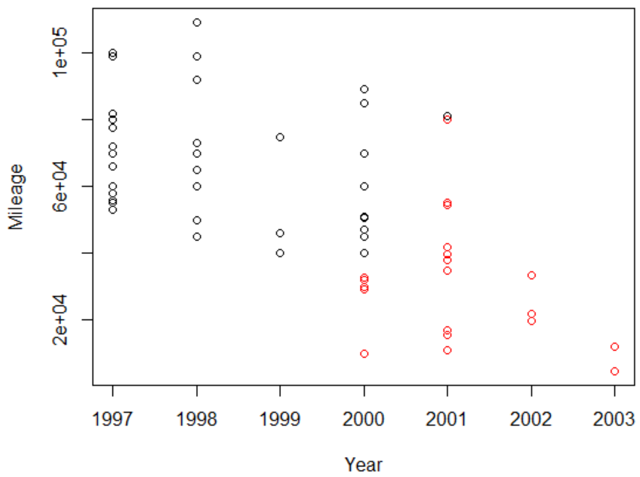
- In R, use glm() with family="binomial"
  - Predicted probabilities: predict(fit,newdata, type="response")

Example Equipment=f(Year, mileage)

Original data



Classified data

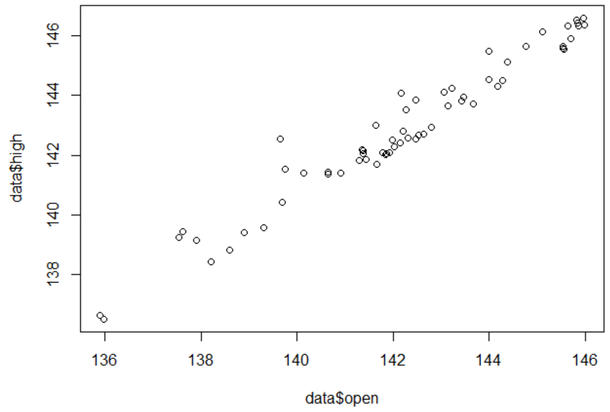


Moving beyond typical distributions

- We know how to model
  - Normally distributed targets -> linear regression
  - Bernoulli and Multinomial targets→logistic regression
  - What if target distribution is more complex?

Example 1: Daily Stock prices NASDAQ
- Open
- High (within day)

Does it seem that the error is normal here?



Example 2: Number of calls to bank
- Y=Number of calls
- X= time

Endless amount of classes → multinomial does not work… (Poisson)

Exponential family

- More advanced error distributions are sometimes needed!
- Many distributions belong to **exponential** family:
  - Normal, Exponential, Gamma, Beta, Chi-squared..
  - Bernoulli, Multinoulli, Poisson…

$$p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})e^{(\boldsymbol{\eta}^T u(\boldsymbol{x}))}$$

- Easy to find MLE and MAP
- Non-exponential family distributions: uniform, Student t

Example: Bernoulli

Generalized linear models

- Assume Y from the exponential family
- **Model** is $Y \sim EF(\mu, \dots),\ f(\mu) = \boldsymbol{w}^T \boldsymbol{x}$
  - Alt $\mu = f^{-1}(\boldsymbol{w}^T \boldsymbol{x})$
  - f^{-1} is activation function
  - f is link function (in principle, arbitrary)
- Arbitrary f will lead to (s – dispersion parameter)

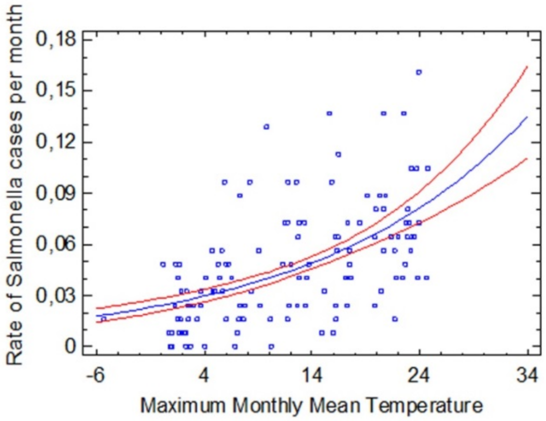
$$p(y|w, s) = h(y, s) g(\boldsymbol{w}, \boldsymbol{x}) e^{\frac{b(\boldsymbol{w}, \boldsymbol{x}) y}{s}}$$

- If f is a canonical link, then

$$p(y|w, s) = h(y, s) g(\boldsymbol{w}, \boldsymbol{x}) e^{\frac{(\boldsymbol{w}^T \boldsymbol{x}) y}{s}}$$

Generalized linear models

- Canonical links are normally used
  - MLE computations simplify
  - MLE $\hat{w} = F(X^T Y)$ → computations do not depend on all data but rather a summary (sufficient statistics)→ computations speed up

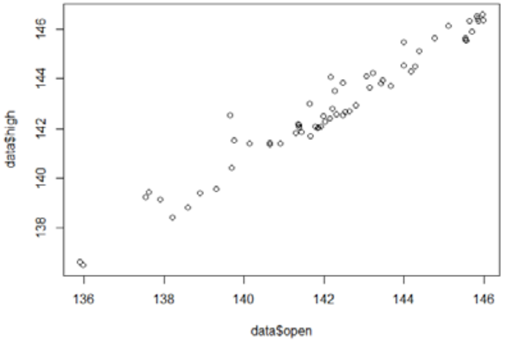


Example: Poisson regression

$$f^{-1}(\mu) = e^{\mu}, Y \sim Poisson\left(e^{\boldsymbol{w}^T \boldsymbol{x}}\right)$$

Generalized linear model: software

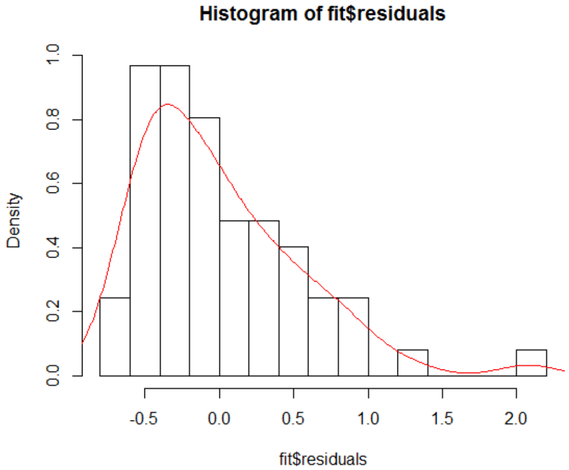
- Use **glm**(formula, family, data) in R



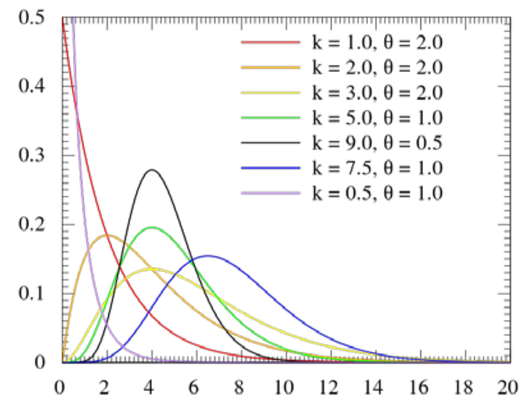
Example: Daily Stock prices NASDAQ

- Open
- High (within day)

1. Try to fit usual linear regression, study histogram of residuals



Gamma distribution: Wikipedia



Generalized linear model

Assume

$$High \sim Gamma(1, \frac{1}{w_0 + w_1 Open})$$

What is link function here?

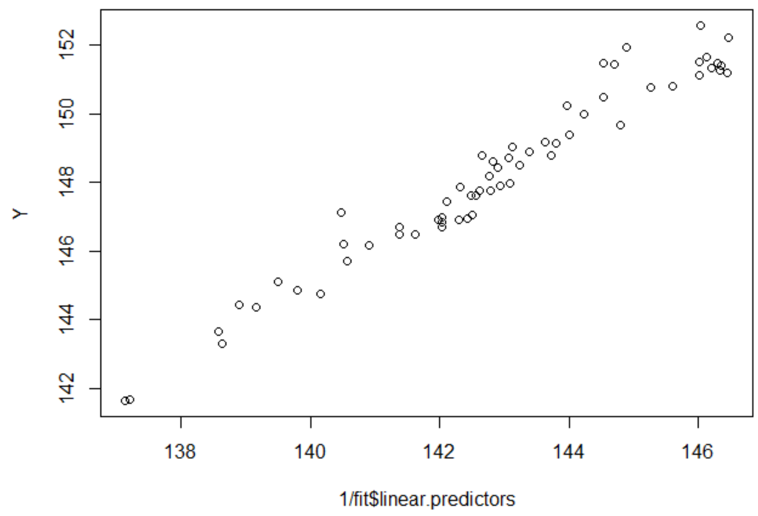
```
Call:
glm(formula = high ~ open, family = Gamma(link = "inverse"),
    data = data)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.0052879 -0.0028896 -0.0006678  0.0016598  0.0148083

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.355e-02  1.962e-04   69.06   <2e-16 ***
open        -4.604e-05  1.379e-06  -33.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

New generated data

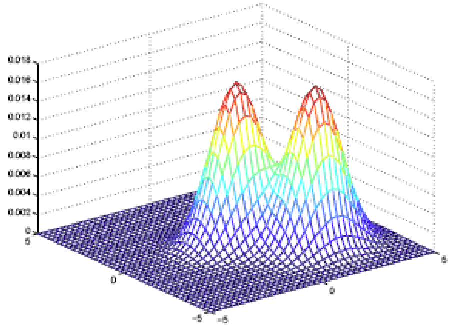
- Has similar pattern as original data!



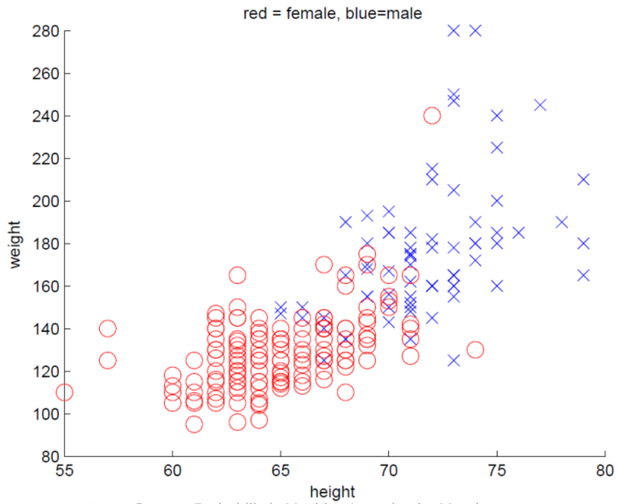
Quadratic discriminant analysis

- Generative classifier
- Main assumptions:
  - $\boldsymbol{x}$ is now **random** as well as y

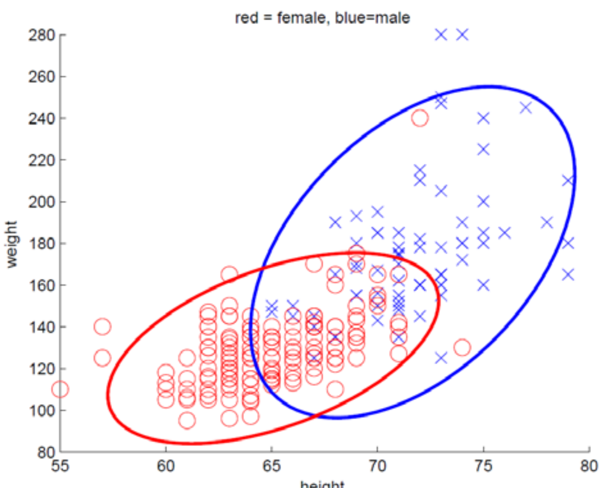
$$p(\boldsymbol{x}|y = C_i, \theta) = N(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$$



Unknown parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}\}$



Source: Probabilistic Machine Learning by Murphy

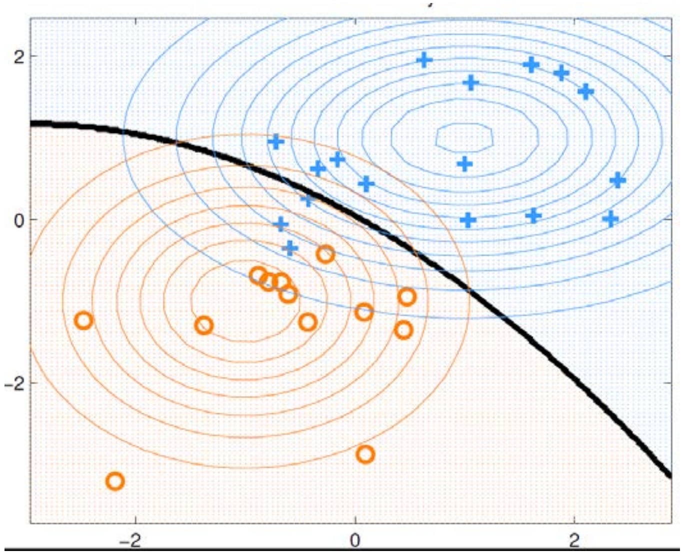


Source: Probabilistic Machine Learning by Murphy

Quadratic discriminant analysis

- If parameters are estimated, classify:

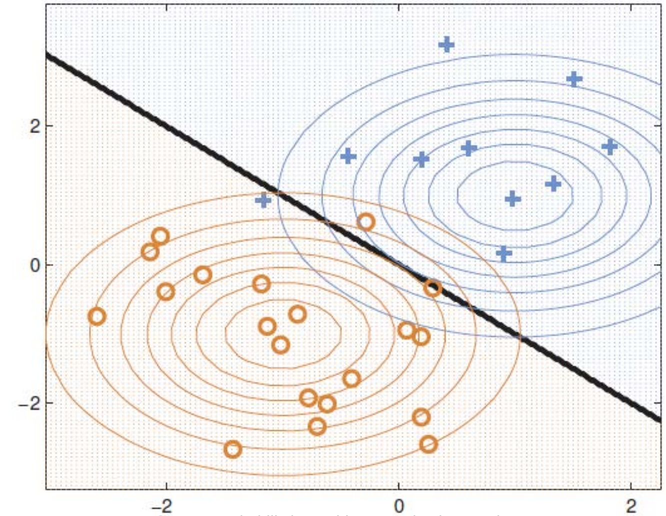
$$\hat{y}(\boldsymbol{x}) = \arg\max_c p(y = c|\boldsymbol{x}, \theta)$$



Source: Probabilistic Machine Learning by Murphy

Linear discriminant analysis (LDA)

- Assumtion $\Sigma_i = \Sigma, i = 1, \dots K$
- Then $p(y = c_i|x) = softmax(w_i^T x + w_{0i})$ →exactly the same form as the logistic regression
  - $w_{0i} = -\frac{1}{2}\boldsymbol{\mu}_i^T \Sigma^{-1} \boldsymbol{\mu}_i + \log \pi_i$
  - $w_i = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i$



Source: Probabilistic Machine Learning by Murphy

- Decision boundaries are linear
  - **Discriminant function**:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

Linear discriminant analysis (LDA)

- Difference LDA vs logistic regression??
  - Coefficients will be estimated differently! (models are different)
- How to estimate coefficients
  - find MLE.

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{N_c} \sum_{i:y_i=c} \mathbf{x}_i, \quad \hat{\Sigma}_c = \frac{1}{N_c} \sum_{i:y_i=c} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_c)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_c)^T$$

$$\hat{\Sigma} = \frac{1}{N} \sum_{c=1}^{k} N_c \, \hat{\Sigma}_c$$

  - Sample mean and sample covariance are MLE!
  - If class priors are parameters (**proportional priors**),

$$\hat{\pi}_c = \frac{N_c}{N}$$

LDA and QDA: code

- Syntax in R, library **MASS**

lda(formula, data, ..., subset, na.action)

- Prior – class probabiliies
- Subset – indices, if training data should be used

qda(formula, data, ..., subset, na.action)

predict(..)

LDA: output

```
resLDA=lda(Equipment~Mileage+Year, data=mydata)
 print(resLDA)
```

```
> print(resLDA)
Call:
lda(Equipment ~ Mileage + Year, data = mydata)

Prior probabilities of groups:
        0         1
0.6440678 0.3559322

Group means:
   Mileage     Year
0 63539.21 1998.447
1 36857.62 2000.762

Coefficients of linear discriminants:
                  LD1
Mileage -1.500069e-05
Year     5.745893e-01
```

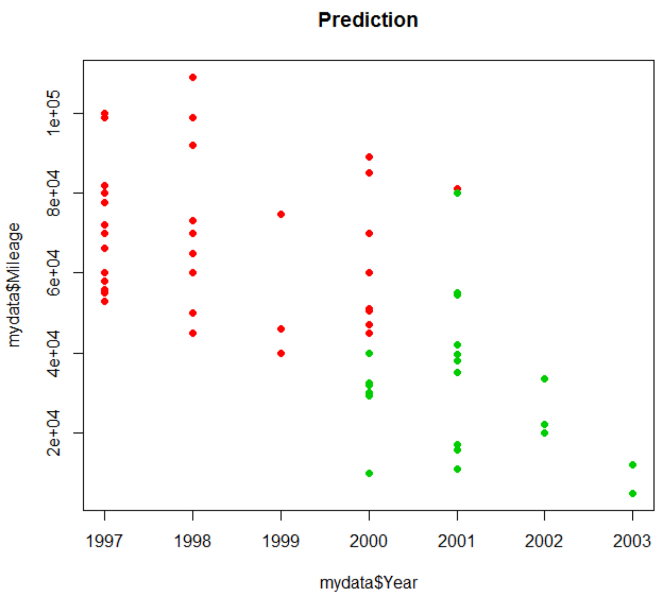
LDA: output

- Misclassified items

```
plot(mydata$Year, mydata$Mileage,
col=as.double(Pred$class)+1, pch=21,
bg=as.double(Pred$class)+1,
main="Prediction")
```

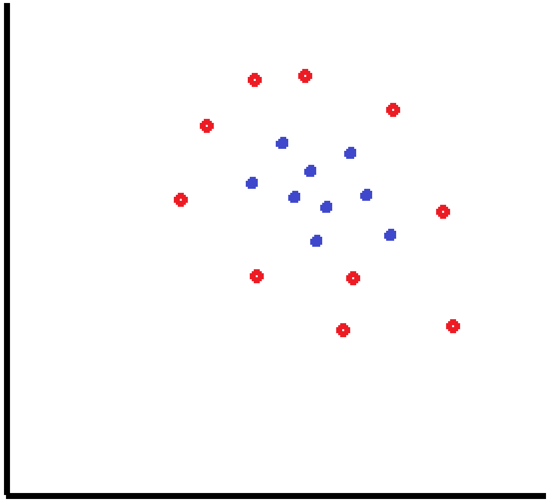
```
> table(Pred$class, mydata$Equipment)

     0  1
  0 31  6
  1  7 15
```



LDA versus Logistic regression

- Generative classifiers are easier to fit, discriminative involve numeric optimization
- LDA and Logistic have same model form but are fit differently
- LDA has stronger assumptions than Logistic, some other generative classifiers lead also to logistic expression
- New class in the data?
  - Logistic: fit model again
  - LDA: estimate new parameters from the new data
- Logistic and LDA: complex data fits badly unless interactions are included



LDA versus Logistic regression

- LDA (and other generative classifiers) handle missing data easier
- Standardization and generated inputs:
  - Not a problem for Logistic
  - May affect the performance of the LDA in a complex way
- Outliers affect $\Sigma \rightarrow$ LDA is not robust to gross outliers
- LDA is often a good classification method even if the assumption of normality and common covariance matrix are not satisfied.