Logistic Regression

B. Ravindran

September 8, 2021

PRML Aug-Nov 2021 (BR section)

Problem Setting

$\chi \subseteq \Re^d$ is the input space

$X = \left(X_1, X_2, \cdots X_d\right)$ is a random variable describing the input

$\Upsilon \subseteq \Re$ or Γ is the output space

Y is a random variable describing the output

$p(X,Y)$ is the data distribution

$p(X,Y) = p(Y|X)p(X)$

$p(Y|x)$ is the predicted output probabilities given an input x

Problem Setting

Instead of assuming that $Y = f(X) + \varepsilon$, one can directly model the $\mathrm{p}\left(G \middle| X\right)$ or $p(Y \middle| X)$

This is sufficient to predict the output labels:

$\hat{G}\left(x\right)$=arg max$_g$ $p\left(Y = g \middle| X = x\right)$

Depending on the assumptions we make on the form of p we get different classifiers.

Simplest of these is the Naive Bayes Assumption

We get LDA assuming that the class conditioned density is Gaussian.

Logistic Regression

Some Notations

- Let $f_k(x)$ be the class conditioned probability, $p(X=x|Y=k)$
- Let π_k be the prior probability of seeing class k, i.e. $p(Y=k)$.
- Then the Bayes theorem can be written as:

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

Recall

- What about multiple classes? $class1 \ldots classK$
- Treat the output variable like any qualitative variable

Let $(x_1, g_1) \ldots (x_N, g_N)$ be the training data. We set

$y_{ik} = 1$ if $g_i = classk$

$= 0$ if $g_i \neq classk$

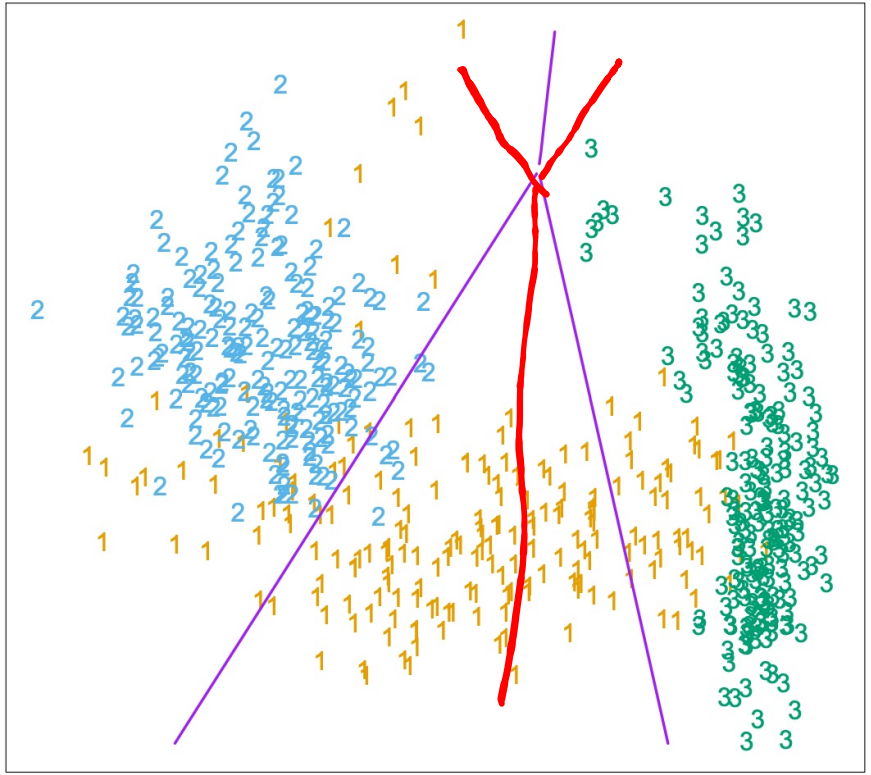
to create a encoded dataset $(x_1, y_1) \ldots (x_N, y_N)$

where each $y_i = y_{i1}, y_{i2}, \ldots y_{iK}$.

Solve for each y_{ik} separately to estimate $\hat{f}_k$

$\hat{G}(x) = \arg\max_{k=1 \ldots K} \hat{f}_k(x)$

Decision Boundaries



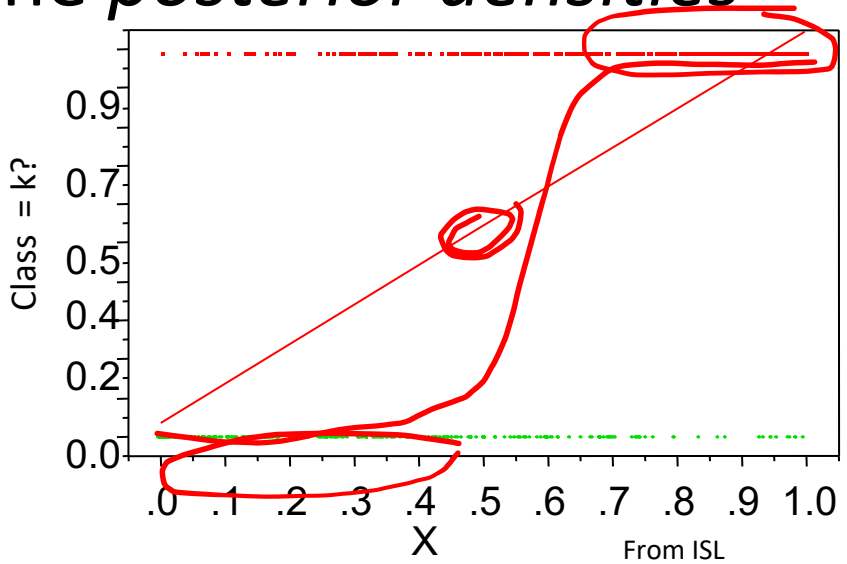
From ESL

Interpreting Linear Regression

- One can consider the $f_k(x)$ fit by linear regression to be the *posterior densities*
- Problem:

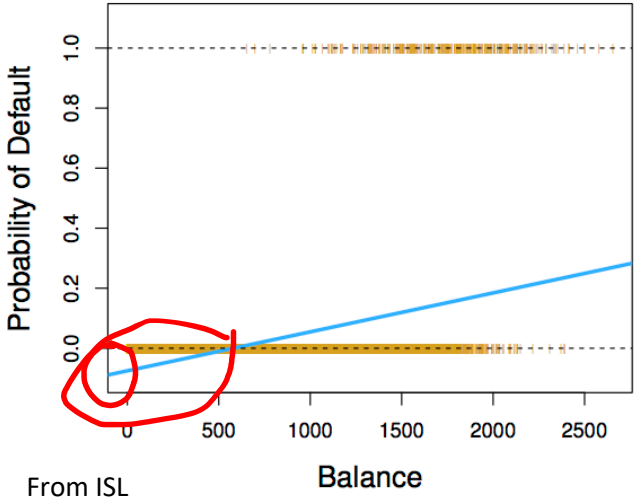
Interpreting Linear Regression

- One can consider the $f_k(x)$ fit by linear regression to be the *posterior densities*
- Problem:



Interpreting Linear Regression

- One can consider the $f_k(x)$ fit by linear regression to be the *posterior densities*
- Problem:



From ISL

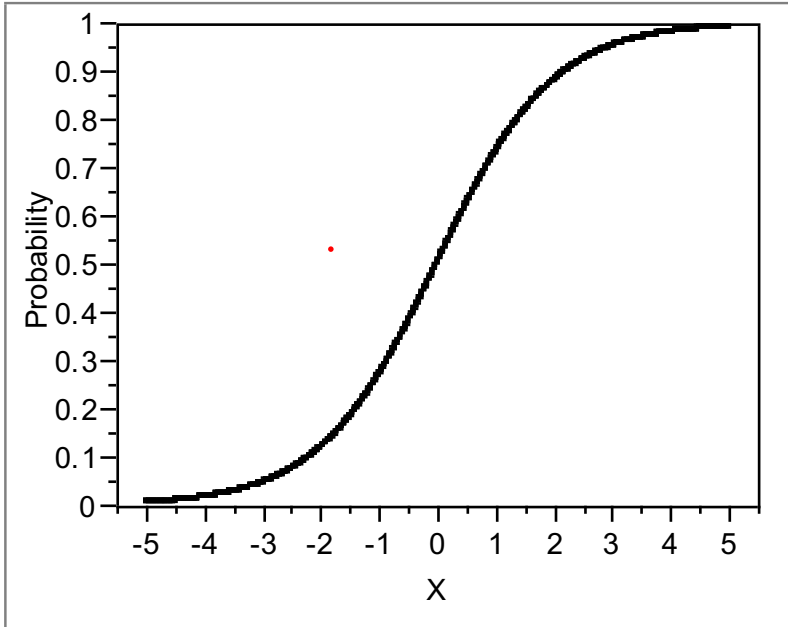
Logistic Function

- Solution: Model the posterior probabilities as a logistic function.

$$P(Y=1|X=x)=\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

$P(Y=0|X=x) = \frac{1}{1+e^{\beta_0+\beta_1 x}}$

From ISL



$$\ln \frac{P(Y=1|x)}{P(Y=0|x)} = \beta_0 + \beta x$$

$$\ln \left\{ e^{\beta_0 + \beta x} \right\} = 0$$

$$\beta_0 + \beta x = 0$$

$$\log\left(\frac{p}{1-p}\right)$$

log odds

logit.

Class boundaries

$$\log \frac{pr(Y=1|X=x)}{pr(Y=K|X=x)} = \beta_{10} + \beta_1^T x$$

$$\log \frac{pr(Y=2|X=x)}{pr(Y=K|X=x)} = \beta_{20} + \beta_2^T x$$

$$\vdots$$

$$\log \frac{pr(Y=K-1|X=x)}{pr(Y=K|X=x)} = \beta_{(K-1)0} + \beta_{K-1}^T x$$

Class boundaries

$$\log \frac{pr(Y=1|X=x)}{pr(Y=K|X=x)} = \beta_{10} + \beta_1^T x$$

$$\log \frac{pr(Y=2|X=x)}{pr(Y=K|X=x)} = \beta_{20} + \beta_2^T x$$

$$\vdots$$

$$\log \frac{pr(Y=K-1|X=x)}{pr(Y=K|X=x)} = \beta_{(K-1)0} + \beta_{K-1}^T x$$

$$P(Y=1|X=x) = \frac{e^{\beta_{10}+\beta_1^T x}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_l^T x}}$$

$$\vdots$$

$$P(Y=k|X=x) = \frac{e^{\beta_{k0}+\beta_k^T x}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_l^T x}}$$

$$\vdots$$

$$P(Y=K|X=x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_l^T x}}$$

Fitting LR models

- Use Maximum Likelihood Estimates
- Assume that your training data is N independent observations
- The likelihood is given by:

$$\ell(\theta) = \sum_{i=1}^{N} \log p_{y_i}(x_i; \beta), \text{ where } \beta \text{ denotes all the parameters}$$

Fitting LR Models

- Assume the 2 class case – easier to derive

Let $p(x)$ denote $p(Y=0 \mid X=x)$

Then $1-p(x)$ denotes the $p(Y=1 \mid X=x)$

[Note that the classes have been coded as 0 and 1]

$$\ell(\beta)=\sum_{i=1}^{N}\left\{y_i \log p(x_i;\beta)+(1-y_i)\log\left(1-p(x_i;\beta)\right)\right\}$$

- Maximize the likelihood w.r.t. beta

Summary

- Logistic Regression is a classification approach!
- Assumes that the class probabilities are given by a logit or sigmoid function
- Directly models the separating surface as a linear function
- Especially popular in binary classification
- Can be combined with Lasso to yield a sparse classifier

Linear Discriminant Analysis

Problem Setting

Instead of assuming that $Y = f(X) + \varepsilon$, one can directly model the $\mathrm{p}\left(G \middle| X\right)$ or $p(Y \middle| X)$

This is sufficient to predict the output labels:

$\hat{G}(x) = \arg\max_g p\left(Y = g \middle| X = x\right)$

Depending on the assumptions we make on the form of p we get different classifiers.

Simplest of these is the Naive Bayes Assumption

Recall, Bayes theorem:

$$p\left(Y \middle| X\right) = \frac{p\left(X, Y\right)}{p\left(X\right)} = \frac{p\left(X \middle| Y\right) p\left(Y\right)}{p\left(X\right)}$$

Some Notations

- Let $f_k(x)$ be the class conditioned probability, $p(X=x|Y=k)$
- Let π_k be the prior probability of seeing class k, i.e. $p(Y=k)$.
- Then the Bayes theorem can be written as:

$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

Linear Discriminant Analysis

- Assume that each class conditioned density is a Gaussian!
  - Further assume that each of the Gaussians differ only in their means, but have identical covariance

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

Linear Discriminant Analysis

- Assume that each class conditioned density is a Gaussian!
  - Further assume that each of the Gaussians differ only in their means, but have identical covariance

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\right)$$

Linear Discriminant Analysis

- Assume that each class conditioned density is a Gaussian!
  - Further assume that each of the Gaussians differ only in their means, but have identical covariance

Predict the output labels:

$$\hat{G}(x) = \arg\max_g p(Y = g | X = x)$$

$$= \arg\max_g \frac{\pi_g f_g(x)}{\sum_{k=1}^{K} \pi_k f_k(x)}$$

LDA

- To train the classifier one has to
  - Estimate the prior class probabilities, π_k
    - Fraction of data points belonging to class k
  - Estimate the class conditioned means, μ_k
    - Average of the data points belonging to class k
  - Estimate the per class variance, Σ
    - Pooled estimate

$$\hat{\Sigma} = \frac{1}{(N-K)} \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \hat{\mu}_k)$$

Class Boundary?

- Compare the probability of two classes
  - Comparing their logarithms sufficient
  - Recall: equal covariance
    - Normalization factors cancel out
    - Quadratic terms in x cancel out

Class Boundary?

$$\log \frac{pr\left(Y=k\middle|X=x\right)}{pr\left(Y=l\middle|X=x\right)} = \log \frac{f_k\left(x\right)}{f_l\left(x\right)} + \log \frac{\pi_k}{\pi_l}$$

$$= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}\left(\mu_k + \mu_l\right)^T \Sigma^{-1}\left(\mu_k - \mu_l\right) + x^T \Sigma^{-1}\left(\mu_k - \mu_l\right)$$

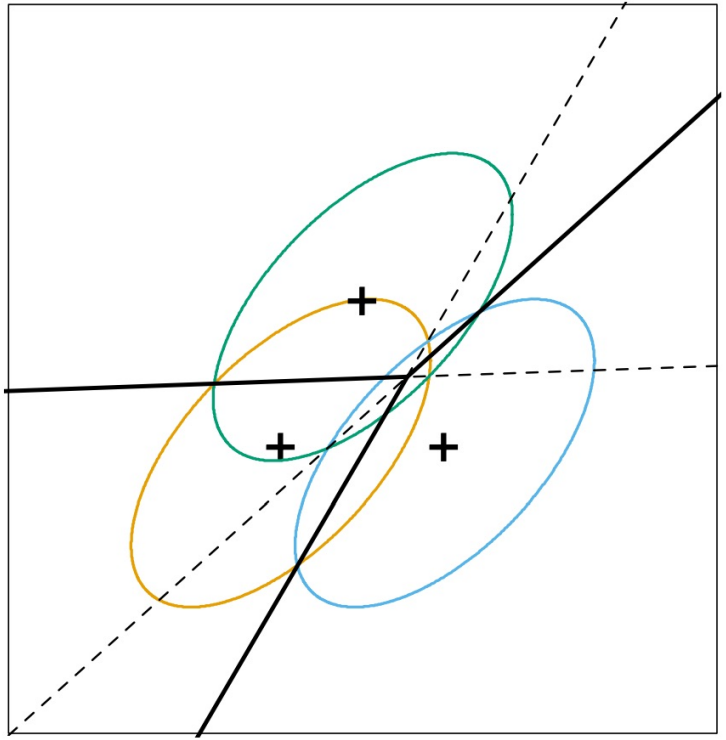
Solve for equal probabilities:

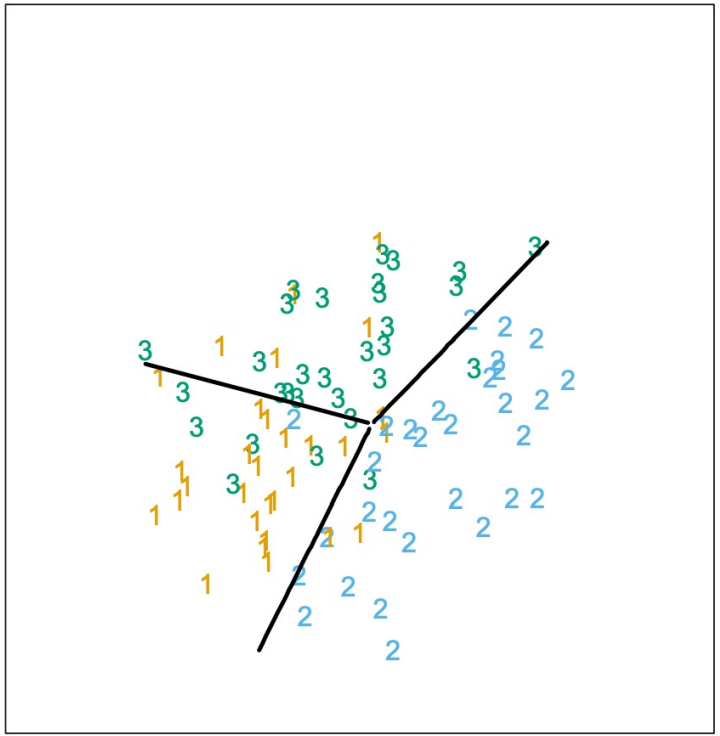
$$pr\left(Y=k\middle|X=x\right) = pr\left(Y=l\middle|X=x\right)$$

$$i.e.,\ \log \frac{pr\left(Y=k\middle|X=x\right)}{pr\left(Y=l\middle|X=x\right)} = 0$$

Hence, $\log \frac{\pi_k}{\pi_l} - \frac{1}{2}\left(\mu_k + \mu_l\right)^T \Sigma^{-1}\left(\mu_k - \mu_l\right) + x^T \Sigma^{-1}\left(\mu_k - \mu_l\right) = 0$

Class Boundaries

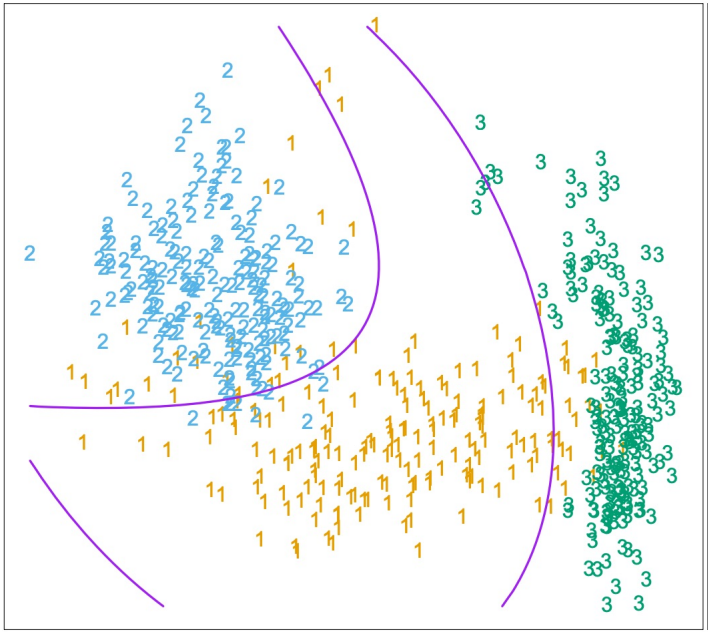




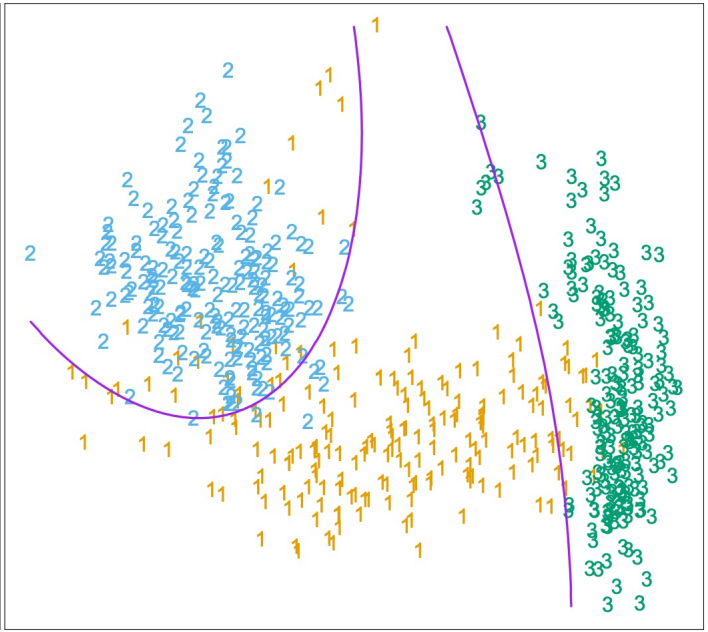
Quadratic Discriminant

- What if the classes do not have the same covariance matrix?
- The quadratic term in the discriminant does not cancel out
- LDA -> QDA
  - Estimate covariance matrices separately

QDA



LDA fit with basis expansion



QDA

Summary

- Assumes the class conditional density is a Gaussian
  - Same covariance for linear
  - Different covariance for quadratic
- Near ideal if the data comes from Gaussian distributions
  - Good approximation even if the assumption is violated
- Can we viewed as a feature selection method also
  - Later lecture

LDA vs LR

- Both produce linear boundaries
- LDA assumes that the observations are drawn from the normal distribution with common variance in each class, while logistic regression does not have this assumption
- Logistic regression is unstable when the classes are well separated
- In the case where N is small, and the distribution of predictors X is approximately normal, then LDA is more stable than Logistic Regression