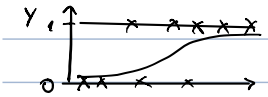
Reminder: Logistic regression

$Y \sim Bernoulli(\pi)$ - indiv. data

$E\{Y\} = \pi, \quad Var\{Y\} = \pi(1-\pi)$

$Y \sim Binomial(n, \pi)$ - grouped data

$E\{Y\} = n\pi \quad Var\{Y\} = n\pi(1-\pi)$



$$\pi = E\{Y\} = g(\beta_0 + \beta_1 X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

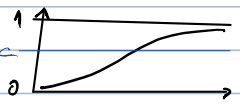
↳ not a linear model

- CDF of logistic distr.

$$Likelihood = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

Logistic regression for classification:

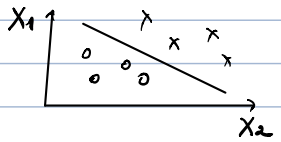
$$\hat{Y} = \begin{cases} 1, & \text{if } \hat{E}\{Y|X\} \geq c \\ 0, & \text{if } \hat{E}\{Y|X\} < c \end{cases}$$



→ confusion matrix
ROC

The classification cutoff separates points by a hyperplane:

$$\frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2)}} = \frac{1}{2} \Rightarrow \beta_0+\beta_1 X_1+\beta_2 X_2 = 0 \Rightarrow X_1 = -\frac{\beta_0}{\beta_1} + \frac{\beta_2}{\beta_1} X_2$$

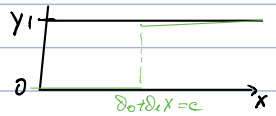


Comments:

- Linear separability: no unique MLE estimates $Var(\beta) \to \infty$
- Alternative ways to define linear decision boundary:

$$f(x) = \begin{cases} 1, & \text{if } X^T\beta \geq c \\ 0, & \text{if } X^T\beta < c \end{cases}$$

} g(x): step function



no prob. interpretation
no MLE

- Variable selection

no residuals ⇒ no R^2, no SSE or MSE

$AIC = -2\log L + 2p$

$BIC = -2\log L + \log N \cdot p$

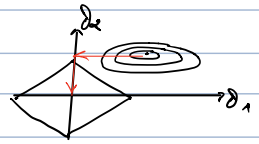
} search for subsets of predictors that minimize AIC or BIC

train/dev/test sets or cross validation

- Regularization: $\min\{-\log L + \lambda \|\beta\|_2^2\}$

components:

(1) objective function, (2) optimization algo



Multi-class classification

possible classes

X	Y = 1	2	...	K
X_1	n_{11}	n_{12}	...	n_{1K}
X_2	n_{21}	n_{22}	...	n_{2K}
	$\pi_1(X)$	$\pi_2(X)$	...	$\pi_K(X)$
X_N	n_{N1}	n_{N2}	...	n_{NK}

grouped data

$n_{ik} \in \{0,1\}$ if indiv. data

multinomial

$$\pi_k(X) = P\{Y = k \mid X = x\}, \quad \sum_{k=1}^{K} \pi_k(X) = 1$$

Multinomial logistic regression:

$$\log \frac{\pi_k(X)}{\pi_1(X)} = \alpha_k + \beta_k X, \quad k = 2, \ldots, K$$

arbitrary category as baseline, to satisfy $\sum_{k=1}^{K} \pi_k = 1$

separate params for each category

} useful if interested in param. interpretation

↳ assumption of multi-class logistic regression

Solving for π_K (in the case of individual data, using $\sum_{k=1}^{K}\pi_k = 1$ and $y_{i1} = 1 - \sum_{k=2}^{K} y_{ik}$)

MLE: $\log \text{Likelihood}_i = \log\left[\prod_{k=1}^{K} \pi_k(x)^{y_{ik}}\right] = \sum_{k=2}^{K} y_{ik} \log \pi_k(x) + \left(1 - \sum_{k=2}^{K} y_{ik}\right) \log \pi_1(x_i)$

for one data point to simplify; for full likelihood, take product over all i

Plug into the likelihood the assumed π_i: $\pi_1(x) = \frac{1}{1 + \sum_{k=1}^{K} e^{\alpha_k + \beta_k x}}$, $\pi_k = \frac{e^{\alpha_k + \beta_k x}}{1 + \sum_{k=1}^{K} e^{\alpha_k + \beta_k x}}$, $k = 2, \dots, K$

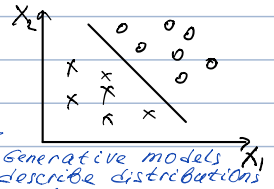
Alternative parametrization: softmax regression

$$\pi_k(x) = \frac{e^{\alpha_k + \beta_k x}}{\sum_{k=1}^{K} e^{\alpha_k + \beta_k x}}$$

more convenient if goal is prediction because symmetric

- only $K-1$ free parameter sets
- set $\alpha_1 = \beta_1 = 0$ to have the previous case

Generative models



Generative models describe distributions of X in each class

Bayes rule: $p(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \propto P(X|Y)P(Y)$

$p(Y|X)$: Linear regression, Logistic regression $\Rightarrow$ discriminative models

$P(X|Y)P(Y)$: generative models

In classification, Y is discrete

$P\{Y = k|X\} = \frac{f_k(x) \cdot \pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\pi_\ell}$

$f_k(x)$: distribution of X in a class $\rightarrow$ likelihoods; π_k: class prior; denominator: normalizing constant

$P\{Y = k|X\}$: most probable class, given the data

Goal: find class k maximizing the posterior probability

MAP (maximum a posteriori learning)

More formally: minimize the 0-1 loss function

$$J(Y, \hat{Y}(x)) = \begin{cases} 0, & \hat{Y}(x) = Y \text{ — correct prediction} \\ 1, & \hat{Y}(x) \neq Y \text{ — wrong prediction} \end{cases}$$

In linear reg., $J(Y, \hat{Y}) = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$

Bayes risk: loss function, averaged over $f_k(x)$ and K:

$$R(\hat{Y}(x)) = E\{E\{J(Y, \hat{Y}(x))\}\}$$

inner E: over $f(X|Y=k)$ - all data of a class; outer E: over π_K - all classes

The MAP estimator minimizing Bayes risk under 0-1 loss is

$$\hat{Y} = \arg\max_k P\{Y = k|X\} = \arg\max_k f_k(X)\pi_k$$

ignore the denominator

Specification of $f_k(x)$ define the classifier:

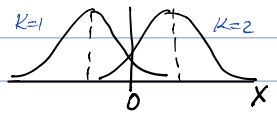
- $f_k(x)$ multivariate Gaussian, same Σ (variance-covariance) per class → LDA (Linear discriminant analysis)
- $f_k(x)$ multivariate Gaussian, different Σ_k → QDA (quadratic discriminant analysis)

$f(x)$ = mixture of Gaussians; $f_k(x)$ – component of the mixture

- $f_k(x) = \prod_{p=1}^{P} f_{kp}(x_p)$ – naïve Bayes
  - in each class, orthogonal dimensions
  - typically (but not always) $f_{kp}(x_p)$ is non-parametric (e.g., Kernel density)

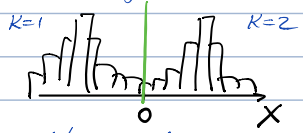
Example: one predictor

Population: $N(\mu_k, \sigma^2)$, same



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

Data: histograms



Q: what is the optimal decision boundary?

$$P\{Y=k|X\} = \frac{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)\cdot\pi_k}{\sum_{l=1}^{2}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)\pi_l}$$

same variance; does not depend on k

Predict class k:

$$\hat{Y}(x) = \arg\max_k P\{Y=k|X\} = \arg\max_k \log P\{Y=k|X\}$$

$$= \arg\max_k \left\{-\frac{1}{2\sigma^2}(x^2 - 2x\mu_k + \mu_k^2) + \log\pi_k\right\}$$

(x^2 does not depend on k)

$$= \arg\max_k \left\{\frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log\pi_k\right\} = \arg\max_k \delta_k(x)$$

$\delta_k(x)$, linear in x → linear discriminant function

Assume $\pi_1 = \pi_2 = \frac{1}{2}$

$$\hat{Y}(x) = \arg\max_k \left\{x\mu_k - \frac{1}{2}\mu_k^2\right\}.$$

The boundary is $\{x\}$ where both classes have same posterior prob:

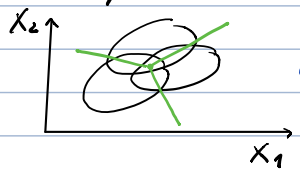
$$x\mu_1 - \frac{1}{2}\mu_1^2 = x\mu_2 - \frac{1}{2}\mu_2^2$$

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

In practice, estimate μ_k, σ and π_k from the data

Multiple classes and predictors:

Population:



- multivariate Gaussian $N(\mu_k, \Sigma)$
  - $p\times 1$, $p\times p$
  - in this case: 2×1, 2×2

$$\hat{Y}(x)_{p\times 1} = \arg\max_k P\{Y=k|X\}$$

Bayes rule

$$= \arg\max_k \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l}$$

$$= \arg\max_k f_k(x)\pi_k$$

$$= \arg\max_k [\log f_k(x) + \log\pi_k]$$

Next, substitute multivariate Normal distribution:

$$\hat{Y}(X) = \arg\max_k \left\{ -\log(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}} - \frac{1}{2}(X-\mu_k)^T \Sigma^{-1}(X-\mu_k) + \log \pi_k \right\}$$

does not depend on k

$(X-\mu_k)^T$: $p\times 1$ $p\times 1$ → $1\times p$; Σ^{-1}: $p\times p$; $(X-\mu_k)$: $p\times 1$; result: 1×1

$$\overset{\text{open}}{=} \arg\max_k \left\{ -\frac{1}{2} X^T\Sigma^{-1}X + X^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log \pi_k \right\}$$

Linear expression in X → $\delta_k(X)$ - linear discriminant function

Decision boundary between classes k and l:

$$\{X: P\{Y=k|X\} = P\{Y=l|X\}\} = \{X: \delta_k(X) = \delta_l(X)\} = \{X: \delta_k(X) - \delta_l(X) = 0\}$$

$$\overset{MVN}{=} \left\{X: X^T\Sigma^{-1}(\mu_k-\mu_l) - \frac{1}{2}(\mu_k-\mu_l)^T\Sigma^{-1}(\mu_k-\mu_l) + \frac{\log \pi_k}{\log \pi_l} = 0\right\}$$

For example, $K=2$:

$$\log\frac{\pi_1}{\pi_2} - \frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) \overset{\text{new notation}}{=} a_0$$

$\Sigma^{-1}(\mu_1-\mu_2)$: $p\times p$, $p\times 1$ → $p\times 1$ $\overset{\text{new notation}}{=} (a_1\ a_2 \dots a_p)^T$ ⇒ classify $\hat{Y}=1$ if $a_0 + \sum_{j=1}^{p} a_j x_j > 0$ → Linear combination of all predictors (i.e., no feature selection!)

Report posterior probability $P\{Y=k|X\} = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l} = \frac{\exp\{-\frac{1}{2}\delta_k(X)\}}{\sum_{l=1}^{K}\exp\{-\frac{1}{2}\delta_l(X)\}}$

Non-linear decision boundaries: create higher order features X_1^2, X_2^2, X_1X_2 etc

<u>Summary</u>: - decision boundary is a linear combination of all features
- no feature selection
- multivariate Gaussian (or at least continuous predictors)