NAIVE BAYES CLASSIFIER

15/3/21

$X = (X_1, \dots, X_p)$

y response (class)

$$P(y = j \mid X = \underline{x}) = \frac{\pi_j \, f_j(\underline{x})}{\sum_{i=1}^{K} \pi_i \, f_i(\underline{x})}$$

$P(y=j)$ $\quad$ $f(\underline{x} \mid j)$

① LDA, $f_j(\underline{x}) \sim N_p(\underline{\mu}_j, \Sigma)$

② QDA, $f_j(\underline{x}) \sim N_p(\underline{\mu}_j, \Sigma_j)$, $j = 1, \dots, k$

③ NB

NAIVE BAYES CLASSIFIER

The main assumption is:

$$f_j(\underline{x}) = \prod_{k=1}^{p} f_{jk}(x_k)$$

$\underline{x} = (x_1, x_2, \ldots, x_p)$

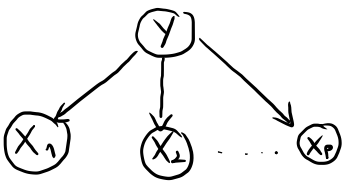
CONDITIONAL INDEPENDENCE

$X_1, \ldots, X_p$ DISCRETE

$$\underbrace{P(X_1 = x_1, \ldots, X_p = x_p \mid Y = j)}_{f_j(\underline{x})} = \underbrace{P(X_1 = x_1 \mid Y = j)}_{f_{j1}(x_1)} \cdot \underbrace{P(X_2 = x_2 \mid Y = j)}_{f_{j2}(x_2)} \cdots \underbrace{P(X_p = x_p \mid Y = j)}_{f_{jp}(x_p)}$$

$X_1, \ldots, X_p$ CONTINUOUS

$$f(x_1, \ldots, x_p \mid j) = f(x_1 \mid j) \cdot \ldots \cdot f(x_p \mid j)$$



Predictors are conditionally independent given the class

(naive, idiots' bayes, ...)

(GAUSSIAN) NAIVE BAYES CLASSIFIER

$f_j(x)$ follows a multivariate Gaussian distribution $\Rightarrow X_1, \dots, X_p$ Gaussian (conditional on class j)

Recall:

(X, Y) normally distribution:

$$X \perp Y \iff \text{Cov}(X, Y) = 0$$

($\perp$: independent)

$\Rightarrow$ always true

$\Leftarrow$ Gaussian

So the matrix

$$\Sigma_j = \begin{pmatrix} \sigma^2_{1j} & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{pj} \end{pmatrix}$$

Σ_j: covariance of $X_1, \dots, X_p$ given $Y = j$

$\sigma^2_{ij} = \text{Var}(X_i \mid Y = j)$

LDA: variances depend on j

QDA: off-diagonal of Σ_j are now zero

NB DECISION SURFACE

$$f_{jk}(x_k) = \frac{1}{\sqrt{2\pi}\,\sigma_{kj}} e^{-\frac{1}{2}\left(\frac{x_k - \mu_{kj}}{\sigma_{kj}}\right)^2}$$

So

$$P(Y = j \mid x_1, \ldots, x_p) \propto \pi_j f_j(\underline{x}) = \pi_j \prod_{k=1}^{p} f_{jk}(x_k)$$

$$= \pi_j \prod_{k=1}^{p} \frac{1}{\sqrt{2\pi}\,\sigma_{kj}} e^{-\frac{1}{2}\left(\frac{x_k - \mu_{kj}}{\sigma_{kj}}\right)^2}$$

Find j which maximises this probability is equivalent to find the j

$$\delta_j(\underline{x}) = \log(\pi_j) - \sum_{k=1}^{p} \log(\sigma_{kj}) - \frac{1}{2}\sum_{k=1}^{p}\left(\frac{x_k - \mu_{kj}}{\sigma_{kj}}\right)^2$$

($\underline{x} \downarrow (x_1, \ldots, x_p)$)

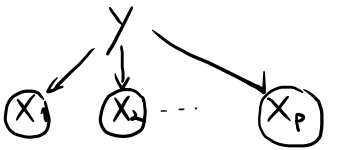
- Quadratic
- Does not involve Σ_j^{-1} so works well for p very large

(MULTINOMIAL) NAIVE BAYES CLASSIFIER

$X_1, \dots, X_p$ categorical

The natural assumption is that f_{jk} is multinomial.

How many parameters ?



Let us assume that Y is binary ($K=2$)

$X_1, \dots, X_p$ are categorical with X_j having G_j categories

$$P(Y=j \mid \underline{x}) \propto P(Y=j)\, P(X_1=x_1 \mid j) \cdots P(X_p=x_p \mid j)$$

$\theta_{y1} = P(Y=0)$
$\theta_{y2} = P(Y=1)$

$\theta_{111} = P(X_1=1 \mid Y=0)$
$\theta_{11G_1} = P(X_1=G_1 \mid Y=0)$
$\theta_{121} = P(X_1=1 \mid Y=1)$
$\theta_{12G_1} = P(X_1=G_1 \mid Y=1)$

$\theta_{p11} = P(X_p=1 \mid Y=0)$
$\theta_{p1G_p} = P(X_p=G_p \mid Y=0)$
$\theta_{p21} = P(X_p=1 \mid Y=1)$
$\theta_{p2G_p} = P(X_p=G_p \mid Y=1)$

X_j $(G_j-1)\times 2$ parameters

θ_{ije} $i=1,\dots,p$ $j=1,2$ $e=1,\dots,G_e$

Local / Global assumption of independence of parameters

PARAMETER ESTIMATION

DATA : $x_1, .., x_n$ → think of $x_1 | Y=0$

independent observations from $f(x; \underline{\theta})$ ← parameters!

① FREQUENTIST :

LIKELIHOOD $\quad L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$

MLE $\quad \hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta)$ $\quad$ maximum likelihood estimate of θ

② BAYESIAN

θ is treated as a random variable with a prior distribution containing information about θ before we see the data [$\theta \sim U(2,3)$]

Combine the likelihood (information from data) and the prior knowledge into a posterior distribution

POSTERIOR DISTRIBUTION

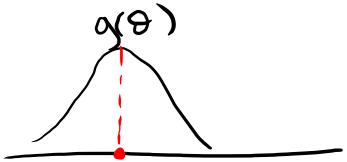
$$g(\theta \mid \text{data}) = \frac{f(\text{data} \mid \theta)\, h(\theta)}{p(\text{data})}$$

($g(\theta \mid \text{data})$: posterior; $h(\theta)$: prior; $f(\text{data} \mid \theta)$: likelihood)

$$= \frac{f(x_1, \theta) \cdot f(x_2, \theta) \cdots f(x_n, \theta)\, h(\theta)}{p(\text{data})}$$

$$\propto L(\theta)\, h(\theta)$$

posterior $\propto$ likelihood $\times$ prior



$$\hat{\theta} = E[\theta \mid \text{data}] = \int \theta\, g(\theta)\, d\theta$$

Bayesian estimate of θ

EXAMPLE (1) BETA-BINOMIAL (BINARY PREDICTOR)

$X \sim \text{Bernoulli}(\theta)$ $\quad f(x;\theta) = \theta^x (1-\theta)^{1-x}$ $\quad x = 0, 1$

θ is between 0 and 1 so a natural prior distribution is

Beta (α, β)

$$h(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad 0 \le \theta \le 1$$

PRIOR (conjugate)

Data: $x_1, \ldots, x_n$

$$L(\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}$$

LIKELIHOOD

$$g(\theta|\text{data}) \propto \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$= \theta^{\sum_{i=1}^{n} x_i + \alpha - 1}(1-\theta)^{n-\sum_{i=1}^{n} x_i + \beta - 1}$$

(α^* and β^*)

POSTERIOR

$$\text{Beta}\left(\sum_{i=1}^{n} x_i + \alpha,\ n - \sum_{i=1}^{n} x_i + \beta\right)$$

updating prior parameters with data

EXAMPLE 2 DIRICHLET - MULTINOMIAL (CATEGORICAL PREDICTOR)

X has G categories

Then the data for n experiments can be summarized as

$$X = (X_1, \dots, X_G)$$

with X_i = "number of times category i is observed out of n trials"

MULTINOMIAL DISTRIBUTION

$$X \sim \text{Multinomial}(\underline{\theta}_1, \dots, \underline{\theta}_G ; n)$$

$$f(\underset{\text{"}}{x} ; \underline{\theta}) = \frac{n!}{\prod_{i=1}^{G} x_i!} \prod_{i=1}^{G} \theta^{x_i}$$

$$(x_1, x_2, \dots, x_G)$$

$$x_i = 0, \dots, n$$

$$\sum_{i=1}^{G} x_i = n$$

$G = 2$ Binomial

ESTIMATION OF $\theta_1, \ldots, \theta_G$

FREQUENTIST

Likelihood

$$L(\underline{\theta}) = \frac{n!}{\prod_{i=1}^{G} x_i!} \prod_{i=1}^{G} \theta_i^{x_i}$$

$$\ell(\underline{\theta}) = \log(n!) - \sum_{i=1}^{G} \log(x_i!) + \sum_{i=1}^{G} x_i \log(\theta_i)$$

MLE

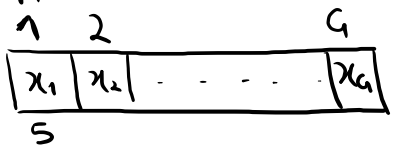
$$\max \ell(\underline{\theta})$$

$$\text{s.t.} \quad \sum_{i=1}^{G} \theta_i = 1$$

$$\longrightarrow \boxed{\hat{\theta}_i = \frac{x_i}{n}}$$

OBSERVED FREQUENCIES

$$\sum_{i=1}^{G} \hat{\theta}_i = \frac{\sum x_i}{n} = 1$$



$n = 100$

BAYESIAN

PRIOR $h(\underline{\theta}) \sim \text{DIRICHLET}(\alpha_1, \dots, \alpha_G)$

$$= \frac{1}{B(\underline{\alpha})} \prod_{i=1}^{G} \theta_i^{\alpha_i - 1} \qquad \text{with } B(\underline{\alpha}) = \frac{\prod_{i=1}^{G} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{G} \alpha_i\right)}$$

$$\sum_{i=1}^{G} \theta_i = 1$$

BETA GENERALIZATION

($G = 2$, $\alpha_1 = \alpha$, $\alpha_2 = \beta$, $\theta_2 = 1 - \theta_1$)

Note: $\alpha_i - 1$ plays the role of x_i in the likelihood, so it can be thought as imaginary counts (before seeing the data)

A common choice is $h(\theta) \sim D(1, 1, \dots, 1)$

$$E(\theta_i) = \frac{\alpha_i}{\sum_{j=1}^{G} \alpha_j}$$

PRIOR FREQUENCIES

POSTERIOR

$$g(\theta \mid \text{data}) \propto L(\underline{\theta})\, h(\underline{\theta}) = \prod_{i=1}^{G} \theta_i^{x_i} \prod_{i=1}^{G} \theta_i^{\alpha_i - 1} = \prod_{i=1}^{G} \theta_i^{x_i + \alpha_i - 1}$$

$$\text{DIRICHLET}(x_1 + \alpha_1, \dots, x_G + \alpha_G)$$

PARAMETER ESTIMATES

$$\hat{\theta}_i^{(B)} = \frac{\alpha_i + x_i}{\sum_{j=1}^{q} \alpha_j + n}$$

BAYESIAN ESTIMATE OF θ_i

$$\hat{\theta}_i^{(MLE)} = \frac{x_i}{n}$$

MAXIMUM LIKELIHOOD ESTIMATE OF θ_i

← n could very small & even zero

Under a uniform prior knowledge:

$$\hat{\theta}_i^{(B)} = \frac{1 + x_i}{q + n}$$

Makes estimation more stable in the case of Naive Bayes classifier (particularly for large p)