Two class problem

$y \in \{0,1\}$ disease, $x \in \{0,1\}$ diagnostic

sample $(x_i, y_i) \overset{iid}{\sim} Pr(x,y)$, $i = 1 \ldots n$

↑ population frequency

$$P(x,y) \equiv \frac{1}{n} \sum_{i=1}^{n} 1(x_i = x)\, 1(y_i = y)$$

↑ sample frequency

Two "models": $P(x \mid y=1)$ vs. $P(x \mid y=0)$

"Bayes factor": $\frac{P(x \mid y=1)}{P(x \mid y=0)} \overset{e.g.}{=} 3.0 \Rightarrow$ "x is 3x more likely under y=1"

i.e. likelihood ratio (LR)

If LR > 1, do we conclude $y=1$?

What about the "base rate" $Pr(y=1)$?

"Prior" model: $P(y=1)$

The correct probabilistic query:

$$P(y=1 \mid x) = \frac{P(y=1 \mid x)\, P(y=1)}{P(x)}$$

"posterior" ← Bayes rule

"likelihood" ↖ ↗ "prior"

↑ "marginal likelihood"

Posterior odds

$$\frac{P(y=1 \mid x)}{P(y=0 \mid x)} = \underbrace{\frac{P(x \mid y=1)}{P(x \mid y=0)}}_{\text{LR}} \times \underbrace{\frac{P(y=1)}{P(y=0)}}_{\text{Prior odds}}$$

e.g. $P(x=1 \mid y=1) = 1.0$ true positive rate of diagnostic

$P(x=1 \mid y=0) = 0.1$ false positive rate of diagnostic

$p(y=1) = 0.001$ rare disease

↳ Posterior odds $= \frac{1.0}{0.1} \times \frac{0.001}{0.999} \approx 0.01$

"data x is 100-to-1 against disease"

Moral: Priors matter! See "Lady drinking tea" (Berger, 1985)

Decision theory: say posterior odds are >1 (e.g. 10).
Do we treat the patient?

decision rule: $\hat{y}(x) \in \{0,1\}$

loss function: $\ell(y, \hat{y}(x))$

For binary classification:

	$\hat{y}=0$	$\hat{y}=1$	
	$\ell(0,0)$ TN	$\ell(0,1)$ FP	$y=0$
	$\ell(1,0)$ FN	$\ell(1,1)$ TP	$y=1$

ℓ_{TN} and ℓ_{TP} should be $\leq \ell_{FN}$ and ℓ_{FP}

but $\ell_{FN} \neq \ell_{FP}$ (not necessarily)

e.g. $\hat{y}(x) = \begin{cases} 1 & \text{if } P(y=1|x) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$

Integrated risk:

$$r(\hat{y}(\cdot)) = \mathbb{E}_{x,y}[\ell(\hat{y}(x), y)]$$

$$= \mathbb{E}_{(x,y)\sim P_r}[\cdots]$$

← population average

Optimal rule

$$\hat{y}^*(\cdot) = \underset{\hat{y}(\cdot)}{\operatorname{argmin}}\ r(\hat{y}(\cdot))$$

$$= \operatorname{argmin}\ \mathbb{E}_x \mathbb{E}_{y|x,x}[\ell(y(x), x)]$$

The optimal rule "always does the right thing":

$$\hat{y}^*(x) = \underset{a}{\operatorname{argmin}}\ \mathbb{E}_{y|x}[\ell(a, x)]$$

"posterior"

In other words we "only" need to know $P_r(y|x)$

↳ $\hat{y}^*(x)$ is the "Bayes decision rule"

↳ $r(\hat{y}^*(\cdot))$ is the "Bayes risk"

Frequentist justification for "Bayesian inference" = posterior computation

---

Modeling: Discriminative vs. generative

We almost never know $P_r(y|x)$ ...

... especially true for higher-dimensional $x \in \mathbb{R}^p$

"making assumptions about" $P_r(y|x)$ vs. $P_r(x|y)P_r(y)$

$X_1 \dots X_p$ are syptoms, $X_j \in \{0, 1\}$ e.g. "is coughing"

$X \equiv X_{1:p}$ takes 2^p possible values

e.g. $p = 40 \Rightarrow 2^{40} \approx 1$ trillion

no hope to estimate $P(y|x)$ $\forall x$ without assumptions

Discriminative example

Assume: $Pr(y|x) = \frac{1}{1 - \exp(-\beta^T x)}$ for some $\beta \in \mathbb{R}^p$

(logistic regression)

Allows us to "share information" across similar X.

Generative example: Naive Bayes

Assume: $Pr(X_{1:p}|y) = \prod_{j=1}^{p} Pr(X_j|y)$

i.e. $X_j \perp\!\!\!\perp X_{j'} | y$ conditionally indep.

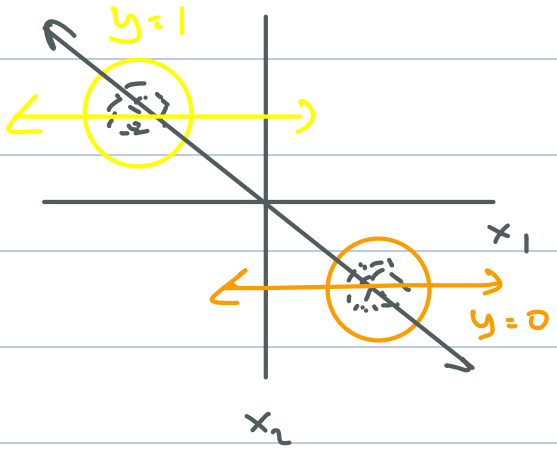
$$P_r(y|x) \overset{\text{Bayes}}{=} \frac{P_r(y)\,P_r(x|y)}{P_r(x)} \overset{\text{naive}}{=} \frac{P_r(y)\prod_j P(x_j|y)}{P_r(x)}$$

what about $P(x) \equiv P_r(x_1, \cdots x_p)$? $\neq \prod_j P_r(x_j)$

$$= \sum_y P_r(y)\,P(x|y) = \sum_y P_r(y)\prod_j P_r(x_j|y)$$

marginalization

Note: $x_j \perp\!\!\!\perp x_{j'} \mid y \not\Rightarrow x_j \perp\!\!\!\perp x_{j'}$



related: Simpson's paradox

Takeaway: Only need to estimate $P(y)$ and $P(x_j|y)$. ↑ $2p$ parameters vs. 2^p

Again, "information sharing".

This model is (often) "wrong" but (often) "useful". See Box.

Data sparsity:

$$P(x_j=1 \mid y=1) \equiv \theta_{j|1} \leftarrow \frac{\sum_i 1(x_{ij}=1,\ y_i=1)}{\sum_i 1(y_i=1)}$$

what if $\theta_{j|1}=1$ (0) ?

e.g. $P(y^{new}=1 \mid x_1^{new}=1, x_2^{new}=1, \cdots)$

$$= \frac{P(y=1)\prod_j \theta_{j|1}}{P(y=1)\prod_j \theta_{j|1} + P(y=0)\prod_j \theta_{j|0}}$$

say $\theta_{1|1} = 1 \quad \theta_{1|0} = 0 \quad \theta_{2|1} = 0 \quad \theta_{2|0} = 1$

$$\text{LR} = \frac{0}{0} \quad \therefore$$

Laplace smoothing

$$\theta_{j|1} \leftarrow \frac{\#(x_j = 1,\ y = 1) + \alpha}{\#(y=1) + 2\alpha}$$

Pierre-Simon Laplace (1700s) proposed this for the "sunrise problem"

See also: Bertrand Russell's "problem of induction" (Bertrand's chicken)

Justification:

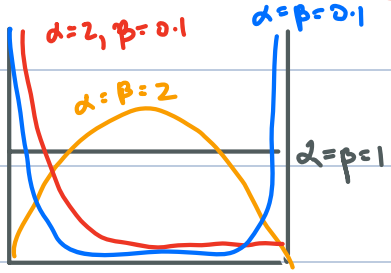
$$x^i \overset{iid}{\sim} \text{Bernoulli}(\theta) \qquad i = 1 \dots n$$

$$\hat{\theta}_{MLE} = \arg\max_\theta \prod_i p(x^i \mid \theta) = \bar{x}$$

Prior: $\theta \sim \text{Beta}(\alpha, \beta)$

"normalizing constant" "kernel"

$$p(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \mathbb{1}(\theta \in (0,1))$$



Alternative parameterization:

$$\mu = \mathbb{E}[\theta \mid \alpha, \beta] = \frac{\alpha}{\alpha+\beta}$$

$$\kappa = \alpha + \beta \leftarrow \text{"concentration"}$$

Posterior:

$$p(\theta \mid x_{1:n}) = \frac{p(\theta)\, p(x_{1:n} \mid \theta)}{p(x_{1:n})} \longrightarrow = \int d\theta\, p(\theta)\, p(x \mid \theta)$$

$$= \int \frac{\text{Beta}(\theta; \alpha, \beta) \prod_i \text{Bern}(x_i; \theta)}{\quad} d\theta$$

$p(x_{1:n})$ is the "normalizing constant" of the posterior

$$\propto_\theta \text{Beta}(\theta; \alpha, \beta) \prod_i \text{Bern}(x_i; \theta)$$

$$= \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \prod_i \theta^{x_i}(1-\theta)^{1-x_i}$$

$$\propto_\theta \theta^{\alpha + \sum_i x_i - 1}(1-\theta)^{\beta + n - \sum_i x_i - 1}$$

This is the kernel of another Beta distribution.

$$\hookrightarrow p(\theta \mid x_{1:n}) = \text{Beta}(\alpha + \textstyle\sum_i x_i,\ \beta + n - \sum_i x_i)$$

No need to "solve" for the normalizing constant explicitly.

Why did this happen? Beta-Bernoulli conjugacy.

Beta is the conjugate prior for the Bernoulli.

Bayesian updating

$\text{Beta}(\alpha + \underbrace{\sum_i x_i}_{\#1s}, \beta + \underbrace{n - \sum_i x_i}_{\#0s})$ α, β are "pseudocounts"

$$\mathbb{E}[\theta \mid x_{1:n}, \alpha, \beta) = \frac{\alpha + \sum_i x_i}{\alpha + \beta + n}$$ ← Laplace smoothing estimator!

$$= \frac{\kappa}{\kappa} \cdot \frac{\alpha}{\kappa + n} + \frac{\sum x_i}{\kappa + n} \cdot \frac{n}{n} = \left(\frac{\kappa}{\kappa + n}\right) u + \left(\frac{n}{n + \kappa}\right) \hat{\theta}_{MLE}$$

$\kappa = \alpha + \beta$ is the strength of the prior / inductive bias.

Subjectivist interpretation of probability

$P(\theta) = \text{Beta}(\theta; \alpha, \beta)$ is not a "frequency"

... it is a "degree of belief."

e.g. P(Bulls win on Thursday)

P(sun comes up tomorrow)

P(Riemann's conjecture is True)

The Bayesian vs frequentist debate was over the interpretation (among other things).