The "True" Bayesian Approach.

MLE $\Leftarrow$ maximum likehood $\quad \hat{\theta}_{MLE} = \arg\max_{\theta} P(D;\theta)$

MAP $\Leftarrow$ max a posteriori $\quad \hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|D)$

---

no single θ captures our belief.

just use $P(\theta|D)$

$$\hat{P}(X,Y) = \int_{\theta} P(X,Y|\theta) \cdot P(\theta|D) \, d\theta$$

$$= \int_{\theta} P(X,Y|\theta) \cdot \frac{P(D|\theta) \cdot P(\theta)}{P(D)} \, d\theta$$

Naive Bayes

$$D = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\} \sim P$$

$$P(D) = P((X_1, Y_1), (X_2, Y_2) \cdots) = \prod_{i=1}^{n} P((X_i, Y_i))$$

MLE estimate of P is $\hat{P}(x,y) = \dfrac{|\{i \mid x_i = x, y_i = y\}|}{n}$

$$\hat{P}(x,y) = \frac{1}{n} \sum_{i=1}^{n} I(x_i = x \text{ and } y_i = y)$$

$I(\text{true}) = 1$, $I(\text{false}) = 0$.

Optimal classifier: $h(x) = \arg\max_{y \in \mathcal{Y}} \hat{P}(x,y)$

$$h(x) = \arg\max_{y \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^{n} I(x_i = x \text{ and } y_i = y)$$

Suppose $\mathcal{X} = \{0,1\}^d$, $\mathcal{Y} = \{-1, 1\}$.

What is the size of the support of $\hat{P}$?

That is, what is $|X \times Y|$? 2^{d+1}

intractable

Naive Bayes Assumption

If I let x_i^α denote the αth feature of the ith example, then

x_i^α and x_i^β are independent given y_i

$x_i^1, x_i^2, \ldots, x_i^d$ are all independent given y_i

$$P(x_i, y_i) = P(x_i|y_i)P(y_i) = \left(\prod_{\alpha=1}^{d} P(x_i^\alpha|y_i)\right)P(y_i)$$

$$= \left[\prod_{\alpha=1}^{d} P(x_i^\alpha|y_i)\right]P(y_i)$$

$$P(D) = \prod_{i=1}^{n}\left(\prod_{\alpha=1}^{d} P(x_i^\alpha|y_i)\right)P(y_i)$$

Examples of tasks where Naive Bayes assumption is reasonable?

- movie recommender system
- text classification (genre/topic)
- demographic info $\Leftarrow$ search
- dating app swipes $\Rightarrow$ secret personal data

Categorical Naive Bayes

$$x \in \{0,1\}^d$$

$$\cancel{P(x^\alpha | y) = \theta}$$

$$P(x^\alpha = 1 | y = c) = \theta_{\alpha c}$$

MLE for $\hat{\theta}_{\alpha c} = \dfrac{|\{i \mid x_i^\alpha = 1,\ y_i = c\}|}{|\{i \mid y_i = c\}|}$

also let $\cancel{\hat{P}(y)}$ $\hat{P}(y=c) = \dfrac{|\{i \mid y_i = c\}|}{n}$

$\{x \mid f(x)\} \equiv$ set of x such that $f(x)$ is true.

$|A|$ = size of set A

$$\frac{\sum_{i=1}^{n} I(x_i^\alpha = 1 \text{ and } y_i = c)}{\sum_{i=1}^{n} I(y_i = c)}$$