The Bayes Optimal Classifier

best of all

Given any probability distribution $\mathcal{D}$ over $\underbrace{\mathcal{X}}_{\text{input space}} \times \underbrace{\{0,1\}}_{\text{binary target}}$, the best label predicting function from $\mathcal{X}$ to $\{0,1\}$ will be

$$f(x) = \begin{cases} \overset{\text{class label}}{1} & \text{if } P[\underbrace{y=1}_{\text{target}} \mid \underline{x}] \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

$$P\{Y=1 \mid \underline{x}\} > P\{Y=0 \mid x\}$$

For every probability distribution $\mathcal{D}$, the Bayes classifier is optimal.

No other classifier has a lower error.

Correct $P[Y \mid \underline{x}]$ is given to us.

Consider the problem of predicting a label $y \in \{0,1\}$ on the basis of a vector of features $\underline{x} = (x_1 \ldots x_d)$ where each x_i is in $\{0,1\}$

Binary feature

The Bayes optimal classifier defines a hypothesis

$$h_{Bayes}(\underline{x}) = \underset{y \in \{0,1\}}{\arg\max}\; P[Y=y \mid X=\underline{x}]$$

Complete posterior

The probability function $P[Y=y \mid X=\underline{x}]$ can be written as a conditional probability table

$((0,0,0,0),0), ((0,0,0,0),1)$

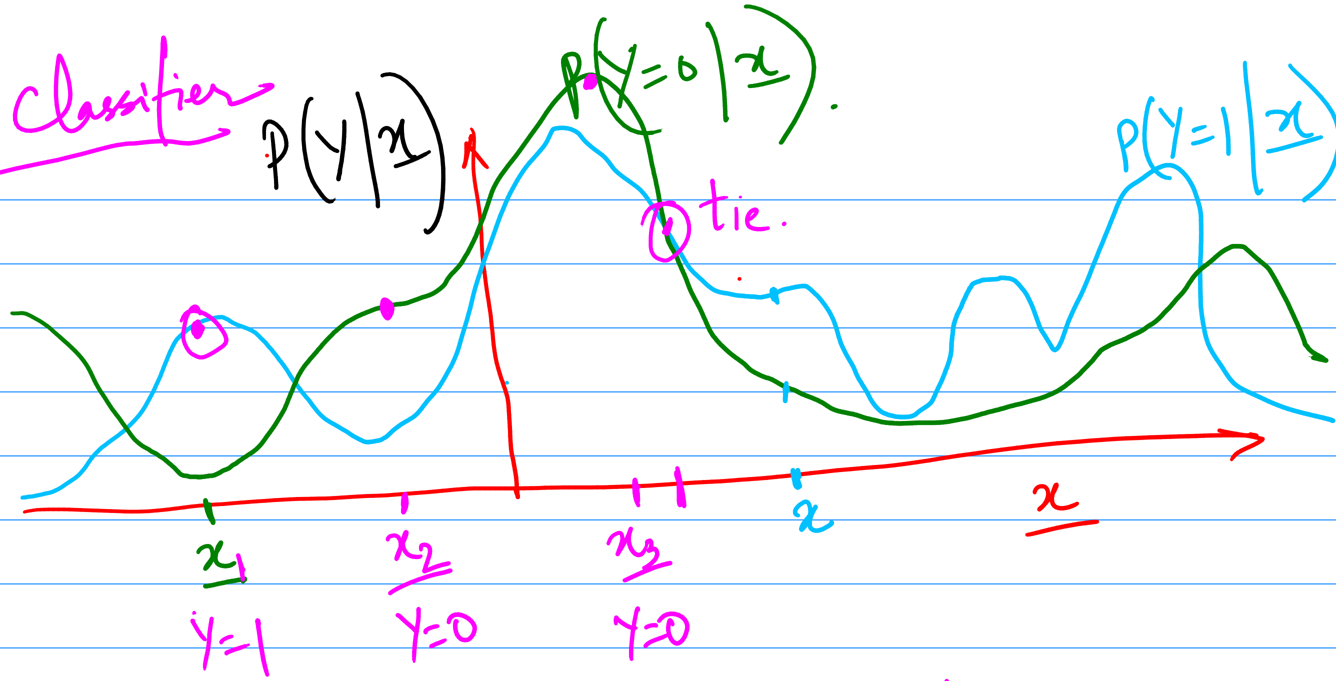
example $\underbrace{X}_{\text{input}} \times \underbrace{Y}_{\text{target}}$

X	x_1 (fixed) x_2 $x_3 \ldots x_d$	$P(Y=1)$	$P(Y=0)$ (implied)
$x_1 \rightarrow$	0 0 0 ⋯ 0	p_1	$1-p_1$
$\underline{x} \rightarrow$	0 0 0 ⋯ 1	p_2	$1-p_2$
$P(Y \mid \underline{x})$	⋮ 2^d feature vectors	⋮	
2^d	1 1 1 ⋯ 1	p_{2^d}	

To store this conditional prob table we need 2^d parameters

$p_1, p_2 \ldots$ governed by the true data generating distr.

Bayes Classifier



$P(Y \mid \underline{x})$ posterior distribution

target. observation

$$\mathbb{E}_{\underline{x}}\left[l(\underline{x}) \right] = \mathbb{E}_{x}\left[1 - P[\underline{x}] \right].$$

Bayes formula.

Prior prob of the label. likelihood

$$P(Y \mid \underline{x}) = \frac{P(Y)\, P(\underline{x} \mid Y)}{P(\underline{x})}$$

Posterior prob

$P(\underline{x}) \rightarrow$ evidence

$$\text{Posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

Estimated through training.

$$\hat{y} = \underset{y}{\arg\max}\; P[Y \mid \underline{x}]$$

Bayes Assumes that the correct Posterior prob is given.

$$= \underset{y}{\arg\max}\; P[Y]\, P[\underline{x} \mid Y]$$

$$= \underset{Y}{\arg\max}\; \frac{P[Y]\, P[\underline{x} \mid Y]}{P[\underline{x}]}$$

$\leftarrow$ does not have Y.

Since the number of parameters grows exponentially with d, the number of examples we need to learn the model also would increase exponentially.

We make a rather naive assumption that given the label, the features are independent of each other.

likelihood
$$\underbrace{P[X = \underline{x}}_{\text{complete feature vector.}} \mid Y = y] = \prod_{i=1}^{d} P[\underbrace{X_i = x_i}_{\text{individual features}} \mid Y = y]$$

Using this naive assumption we can simplify the Bayes hypothesis

$$h_{Bayes}(\underline{x}) = \underset{y \in \{0,1\}}{\operatorname{argmax}} \; \underbrace{P[Y = y \mid X = \underline{x}]}_{\text{posterior.}}$$

$$= \underset{y \in \{0,1\}}{\operatorname{argmax}} \; \underbrace{P[Y = y]}_{\text{prior}} \; \underbrace{P[X = \underline{x} \mid Y = y]}_{\text{likelihood.}}$$

Naive Bayes Classifier
$$= \underset{y \in \{0,1\}}{\operatorname{argmax}} \; \underbrace{P[Y = y]}_{1.} \; \prod_{i=1}^{d} \underbrace{P[X_i = x_i \mid Y = y]}_{2d \text{ per feature likelihood}}$$

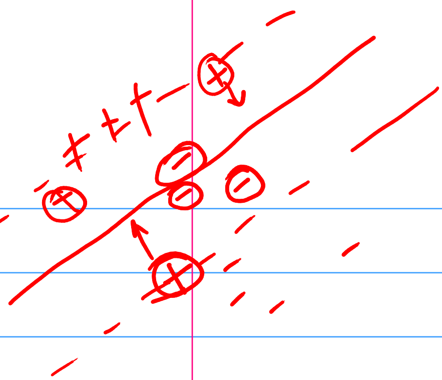
The number of parameters we need to estimate now is

$\underline{2d + 1.}$

$Y=$	$P(x_i=0)$	$P(x_i=1)$
0	1	
1	2.	

This is the Naive Bayes Classifier.

The parameters are estimated using the maximum likelihood principle.



Spam filtering.

Mail. → features → NBC → Y

2^{100}

$x_1 \quad x_2 \quad x_3 \quad x_4 \quad \dots \quad x_{100}$ | $Y = 0, 1$. spam
Not spam

NBC

$x_1 = \quad v_1 \quad v_2 \quad v_3$.

Treating features as independent given the label

$P(x_1 = v_1 \mid Y = 0)$ ← $\dfrac{P(x_1 = v_1 \mid Y = 0)}{P(Y=0)}$ Not spam $Y=0$

$P(x_1 = v_2 \mid Y = 0)$ ←

$P(x_1 = v_3 \mid Y = 0)$ ←

$\dfrac{\text{count}(x_1 = v_1)}{\# \text{total}}$. Not spam $Y = 0$

Gaussian

Mixture of Gaussians

Multivariate Gaussian distribution. x

continuous x — density function $p(x)$