Bayes Classifiers

Training time

Joint model $p(y, \underline{x}) = p(y)\,p(\underline{x} \mid y)$

$p(y=k) = \pi_k \approx \frac{\#\ k \text{ labels}}{N}$

$p(\underline{x} \mid y=k)$ ... eg $N(\underline{x}; \mu^{(k)}, \Sigma^{(k)})$

Not Bayesian — Assuming we know all parameters.

$\mu^{(k)}, \Sigma^{(k)}$ = Mean & cov of $\underline{x}$'s in class k

Naive Bayes $p(\underline{x} \mid y=k) = \prod_d p(x_d \mid y=k)$

Univariate Gaussian, Discrete, ...

Test time

$p(y \mid \underline{x}) \propto p(y, \underline{x})$ (Bayes' Rule)

$y_{guess} = \underset{k}{\operatorname{argmax}}\ p(y=k, \underline{x})$

"goodness"

$N(\underline{x}; \mu^{(0)}, \Sigma^{(0)})$

$y=0$

quadric decision boundary.

$N(\underline{x}; \mu^{(1)}, \Sigma^{(1)})$

$y=1$

Decision boundary

$\pi_0 N(\underline{x}; \mu^{(0)}, \Sigma^{(0)}) =$

$\pi_1 N(\underline{x}; \mu^{(1)}, \Sigma^{(1)})$

$$p(y|\underline{x}) = \frac{p(y, \underline{x})}{p(\underline{x})}$$

$$p(\underline{x}) = \sum_{y'} p(y', \underline{x})$$

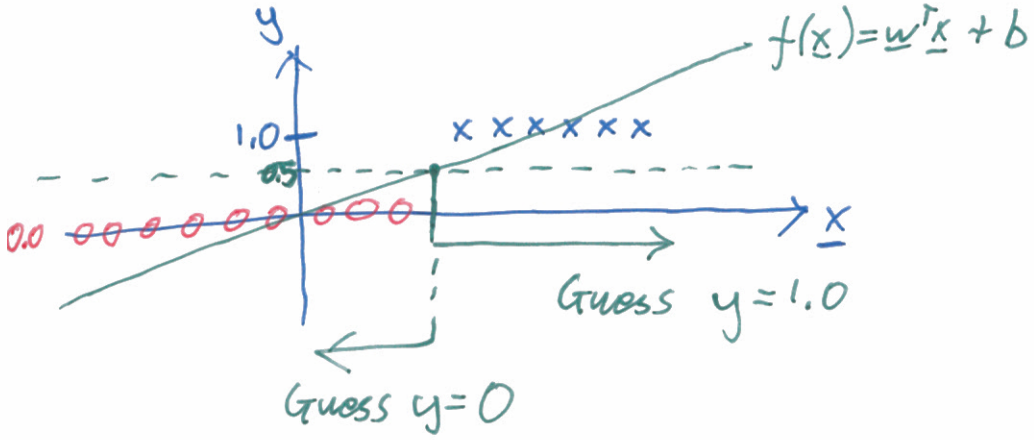
Discipuli Domini Colini Drummond qui vigesimo-septimo die Februarij MDCCXIX subscripserunt. 1719

Arch Rennie 4
Wm Aitkine 2
Geo: Gordon 2
Geo: Paxton 4
Gul: Horsburgh 3
Hen: Ker 4
Joan: Boston 3
Jo: Carruthers 4
Joan Morison 3
John Paxton 2.
Jo Faull 2
Mich: Robertsone 2
Pat Murdoch 3
Simon Elliot 2
Thomas Carmichael 2

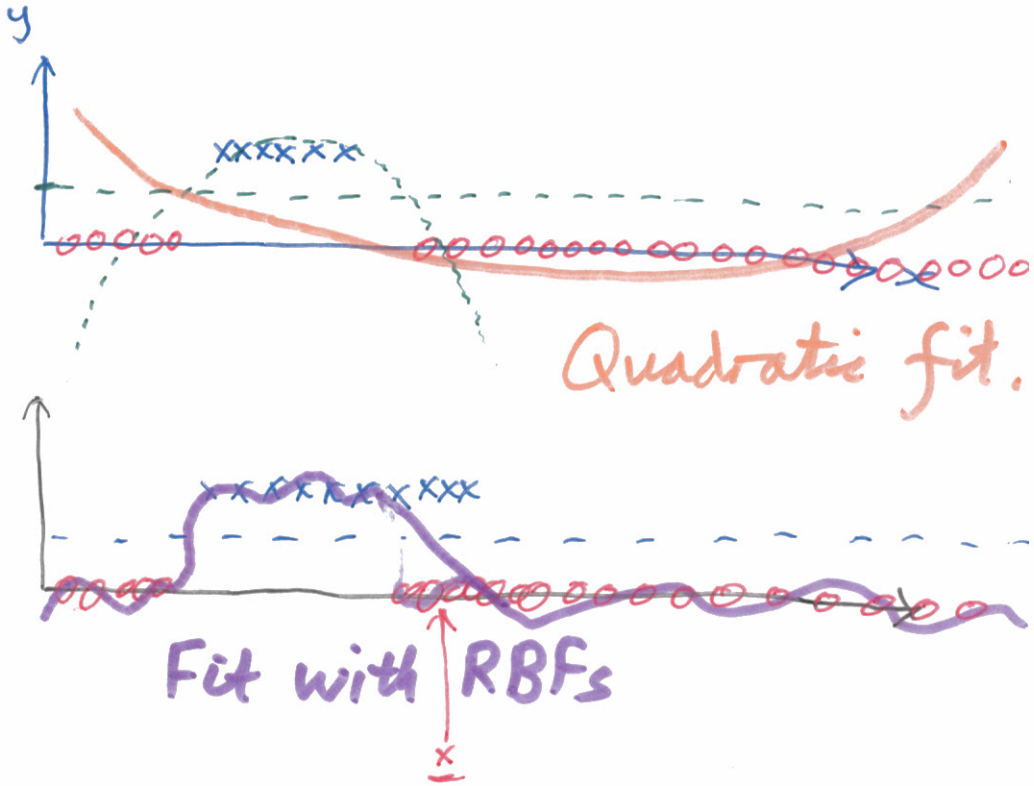
Jam: Bonar 3
Geo: Carruthers 4
Geo: Dewar
Geo: M'Lerie 2
Gul: Ramsay 3
Gul: Dick 4
John Connell 2
Jn Gilchrist mer 2
Jo Wilson
Jo Shiel 1
Jos Jaques 1
Rich: Bell
Ro: Dalrymple 1
Rou: Dunbar 4
Skyrme Smith
Tho: Davidson 2

Alexr Crokat 3
David Lindsay 1
Geo. Doag 4 2
Gul: Taylor 1
Jo: Bardey 2
Jo: Blackhurst
Jo: Horsley
Jotan: Patoun 1
John Rutherfurd 2
Jo: Smith 1
Jo: Thomson 2
Isa. Maddox
Rob: Cleland 4
Rob: Douglass 1
Rob: Richarby 1
Th. Bayes
Tho: Morison 2

Regression for classification



If $f(\underline{x}) > 1/2$, guess $y = 1$



If minimize square loss?

$$\min \underset{p(y|\underline{x})}{\mathbb{E}}\left[(y-f(\underline{x}))^2\right] \quad \text{at some } \underline{x}$$

Cost

$$= p_1(1-f)^2 + \underbrace{(1-p_1)}_{p(y=0|\underline{x})}(0-f)^2$$

$$\uparrow p(y=1|\underline{x})$$

$$= p_1(1-2f+f^2) + (1-p_1)f^2$$

$$= f^2(\cancel{p_1 - p_1} + 1) - 2p_1 f + p_1$$

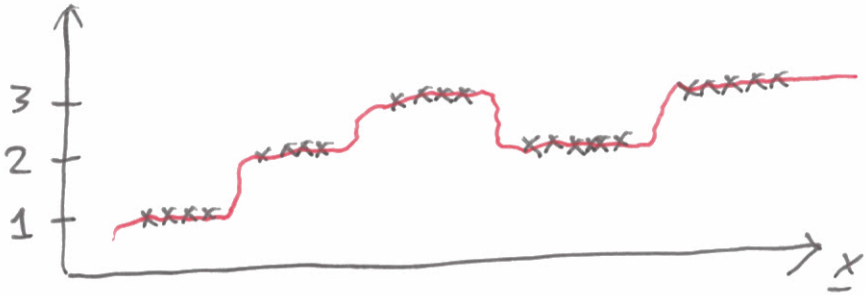
$$\frac{\partial \text{cost}}{\partial f} = 2f - 2p_1 = 0 \quad \text{at optimum}$$

$$\boxed{f = p_1}$$

Multiple classes

$y \in \{1, 2, 3, \ldots 103$

"sport" "crime" "romance"



$f(\underline{x}) = \underline{w}^T \underline{x}$ [Replace $\underline{x}$ with $\phi(\underline{x})$ if you like]

$f(\underline{x}^{(1)}) \approx 1 \Rightarrow$ "sport"

$f(\underline{x}^{(2)}) \approx 3 \Rightarrow$ "romance"

$f\left(\frac{\underline{x}^{(1)} + \underline{x}^{(2)}}{2}\right) \approx 2 \Rightarrow$ "crime"

One-hot encoding, One-of-k-encoding

Vector output

$$\underset{k\times 1}{\underline{y}^{(n)}} = [0\ 0\ 0 \ldots 0\ 0\ \underset{k^{th}\text{ position}}{1}\ 0\ 0 \cdots 0]^T$$

for k classes.

If n^{th} example is in class k

Fit k functions, one for each bit y_k

Predict class where $f_k(\underline{x})$ biggest

Pre-processing also useful for features $\underline{x}$

$x_d \in \{$"red", "green", "blue"$\}$

$\{1, 2, 3\}$

3 features

red → 1 0 0

green → 0 1 0

blue → 0 0 1

1 star 1 0 0 0

2 star 1 1 0 0

3 star 1 1 1 0

Puzzle: R libraries don't create. Why?