Bayes Classifiers

Training time

Joint model \( p(y, x) = p(y) p(x | y) \)

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

\( p(x | y = k) \) ... eg \( N(x; \mu(k), \Sigma(k)) \)

Not Bayesian

Assuming we know all parameters.

Naive Bayes \( p(x | y = k) = \prod_d p(x_d | y = k) \)

Test time

\( p(y | x) \propto p(y, x) \) (Bayes' Rule)

\( y_{\text{guess}} = \arg \max_k p(y = k, x) \)

"goodness"
\[
N(x; \mu^{(i)}, \Sigma^{(i)})
\]
\[y = 0\]

\[
N(x; \mu^{(*)}, \Sigma^{(*)})
\]
\[y = 1\]

Decision boundary

\[
\pi_0 N(x; \mu^{(i)}, \Sigma^{(i)}) = \pi_1 N(x; \mu^{(*)}, \Sigma^{(*)})
\]

\[
p(y|x) = \frac{p(y, x)}{p(x)}
\]

\[
p(x) = \sum_y p(y, x)
\]
Discipuli Domini Colini Drummond qui vigesimo-septimo die Februarii MDCXXIX subscripti sunt

Arch: Rennie 4
Geo. Nicolson 2
Geo: Horsburgh 2
Ken: Ker 1
Jordan: Boston 2
Jo: Carruthers 1
Joan: Morison 3
John: Paxton 2
Jo: Teall 2
Mick: Robertson 2
Pat: Murdoch 3
Simon: Elliot 2

Thomas Carmichael 2
Th: Davidson 2

Alex: Crochat 3
David: Lindsay 1
Geo: Tous 1
Geo: Taylor 1
Jo: Barlow 2
Jo: Ritchie 1
Jo: Horsley

Joan: Patoun 1
John: Adair 1
Jo: Smith 1
Jo: Thomson 2
Isa: Maddox
Rob: Cleland 1
Rob: Douglas 1
Rob: Richardson 1
Th: Bayes
Th: Morison 2
Regression for classification

$f(x) = w^T x + b$

Guess $y = 1.0$

Guess $y = 0$

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

Quadratic fit.

Fit with RBFs.
If minimize square loss?

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

Cost

$$= \pi (1 - f)^2 + (1 - \pi)(0 - f)^2$$

$$\uparrow p(y=1|x) \quad \downarrow p(y=0|x)$$

$$= \pi (1 - 2f + f^2) + (1 - \pi)f^2$$

$$= f^2 (\pi - \pi(0) + 1) - 2\pi f + \pi$$

$$\frac{\text{d} \text{cost}}{\text{d}f} = 2f - 2\pi f = 0 \text{ at optimum}$$

$$f = \pi$$
Multiple classes

\[ y \in \{1, 2, 3, \ldots, 103\} \]

"sport" \quad "romance" \quad "crime"

\[ f(x) = w^T x \]

[Replace \( x \) with \( \phi(x) \)]

\[ f(x^{(1)}) \approx 1 \Rightarrow \text{"sport" if you like} \]

\[ f(x^{(2)}) \approx 3 \Rightarrow \text{"romance"} \]

\[ f\left(\frac{x^{(1)} + x^{(2)}}{2}\right) \approx 2 \Rightarrow \text{"crime"} \]
One-hot encoding, One-of-k-encoding

Vector output

\[ y^{(n)} = [0 \ 0 \ldots \ 0 \ 1 \ 0 \ldots \ 0]^T \]

For \( k \) classes. If \( n \)th example is in class \( k \)

Fit \( k \) functions, one for each bit \( y_k \)

Predict class where \( f_k(x) \) biggest

Pre-processing also useful for features \( x \)

\[ x_2 \in \{ \text{"red", "green", "blue"} \} \]

\[ \epsilon \{ 1, 2, 3 \} \]

3 features

\[ \begin{align*}
\text{red} & \rightarrow 100 \\
\text{green} & \rightarrow 010 \\
\text{blue} & \rightarrow 001
\end{align*} \]

Puzzle: R libraries don't create. Why?