Naïve Bayes

Donald E. Brown

School of Data Science
University of Virginia
Charlottesville, VA 22904

Why use Naïve Bayes

- Assumes features of a multidimensional likelihood are independent even when they are not
- Suppose likelihood is $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of dimension M
- Parameter estimation is $O(\frac{M^2}{2})$
- Complexity of parameter estimation can be worse with other joint distributions
- Suboptimal approaches with good parameter estimates are better than optimal approaches with bad parameter estimates

Naïve Bayes Classifier

- Approximate $f(\boldsymbol{x}|g_i)$ for $i = 1, \ldots, p$ using an independence assumption

$$f(\boldsymbol{x}|g_i) = \prod_{k=1}^{p} f(x_k|g_i)$$

- Estimate or assume the form of $f(x_k|g_i)$
  - Assume each of the class conditional densities is Gaussian.
  - Use kernel or mixture model estimates for $f(x_k|g_i)$.
  - Categorical variables with Bernoulli or mulitnomial estimates.

Naïve Bayes, Gaussian

For $i \in \{1, \ldots, K\}, j \in \{1, \ldots, p\}, \boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$

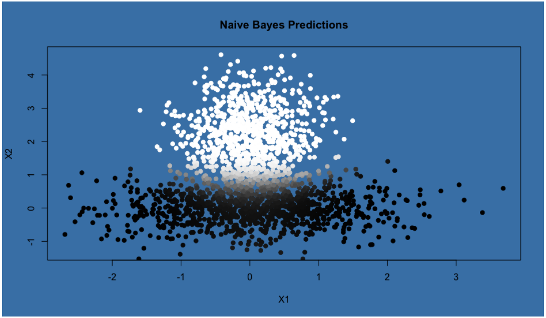
$$
\begin{aligned}
f(\boldsymbol{x}|g_i) &\sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \\
&= \prod_{j=1}^{p} \mathcal{N}(\mu_{ji}, \sigma_{ji})
\end{aligned}
$$

For $i, k \in \{1, \ldots, K\}$ choose $G = g_i$ if

$$
\begin{aligned}
\frac{f(\boldsymbol{x}|g_i)}{f(\boldsymbol{x}|g_k)} &> \frac{Pr[G = g_k]}{Pr[G = g_i]} \\
\frac{\prod_{j=1}^{p} \mathcal{N}(\mu_{ji}, \sigma_{ji})}{\prod_{j=1}^{p} \mathcal{N}(\mu_{jk}, \sigma_{jk})} &> \frac{Pr[G = g_k]}{Pr[G = g_i]}
\end{aligned}
$$

BTA

Example Predictions



Naïve Bayes, Bernoulli

For $i \in \{1, \ldots, K\}, j \in \{1, \ldots, p\}, \theta_{ji} \in (0, 1), \boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$

$$f(\boldsymbol{x}_h|g_i) = \prod_{j=1}^{p} \theta_{ji}^{x_{hj}} (1 - \theta_{ji})^{(1-x_{hj})}, h = 1, \ldots, N$$

For $i, k \in \{1, \ldots, K\}$ choose $G = g_i$ if

$$\frac{f(\boldsymbol{x}|g_i)}{f(\boldsymbol{x}|g_k)} > \frac{Pr[G = g_k]}{Pr[G = g_i]}$$

$$\frac{\prod_{j=1}^{p} \theta_{ji}^{x_j} (1 - \theta_{ji})^{(1-x_j)}}{\prod_{j=1}^{p} \theta_{jk}^{x_j} (1 - \theta_{jk})^{(1-x_j)}} > \frac{Pr[G = g_k]}{Pr[G = g_i]}$$

BTA

Example - Text Classification

- Suppose we have a corpus, $\mathcal{D}$, of documents and we want to classify them into class labels $c \in \mathcal{C}$
- Examples:
  - $\mathcal{D}$ may be reviews and $\mathcal{C} = \{\text{positive}, \text{neutral}, \text{negative}\}$
  - $\mathcal{D}$ may be email and $\mathcal{C} = \{\text{ham}, \text{spam}\}$
- Bag of words, so $\mathcal{W}$ is the vocabulary and $\mathcal{W}_j$ are the words in document $j, j = 1, \ldots, |\mathcal{D}|$
- Bayes optimal classifier: for some document d

$$c^* = \underset{c_i \in \mathcal{C}}{\text{argmax}}\ P(\mathcal{W}_d | c_i) P(c_i))$$

Example - Naïve Bayes Likelihood Estimation

- Likelihood - Independence assumption

$$P(\mathcal{W}_d|c_i)P(c_i) = \prod_{w_j \in \mathcal{W}_d} P(w_j|c_i)P(c_i)$$

$$\hat{P}(w_j|c_i) = \frac{\sharp\,(w_j, c_i)}{\sum_{w \in \mathcal{W}} \sharp\,(w, c_i)}$$

- A word, w_j associated with a class may not appear in the training set, so $P(w_j|c_i) = 0$, which makes the likelihood zero.
- Naïve Bayes smoothing: add $\alpha > 0$

$$\hat{P}(w_j|c_i) = \frac{\sharp\,(w_j, c_i) + \alpha}{\sum_{w \in \mathcal{W}}(\sharp\,(w, c_i) + \alpha)} = \frac{\sharp\,(w_j, c_i) + \alpha}{\sum_{w \in \mathcal{W}} \sharp\,(w, c_i) + \alpha|\mathcal{W}|}$$

Example - Supervised Learning with Naïve Bayes

- Obtain a training corpus of restaurant reviews, $\mathcal{D}_T$, with labels from $\mathcal{C}$
- Priors: Let $d(c_i) \in \mathcal{D}_T$ be a document of class, c_i

$$P(c_i) = \frac{\sharp\left(d(c_i)\right)}{|\mathcal{D}_{\mathcal{T}}|}$$

- Likelihoods: Let $\mathcal{W}_{d(c_i)}$ be words in all $d(c_i) \in \mathcal{D}_T$ and $v = |\mathcal{W}_{d(c_i)}|$

$$\hat{P}(w_j|c_i) = \frac{\sharp\left(w_j \in \mathcal{W}_{d(c_i)}\right) + \alpha}{\sum_{k=1}^{v} \sharp\left(w_k \in \mathcal{W}_{d(c_i)}\right) + \alpha v}$$

Example Words in Reviews

Words	$\sharp\,(w_j\|\text{Positive})$	$\sharp\,(w_j\|\text{Neutral})$	$\sharp\,(w_j\|\text{Negative})$
we	1254	612	1478
I	1090	312	856
you	347	121	538
they	1688	976	2005
horrible	0	2	883
bad	362	439	3795
good	2183	729	691
liked	2847	837	114
tasty	884	33	17
salmon	158	26	39
tuna	112	15	137
calamari	2	0	0

Let $\alpha = 1.5$

Restaurant Review Posteriors

Text	$f(w_j \vert \text{Positive})$	$f(w_j \vert \text{Neutral})$	$f(w_j \vert \text{Negative})$
I	1.09E-2	3.13E-3	8.57E-3
liked	2.84E-2	8.38E-3	1.15E-3
the	NA	NA	NA
tasty	8.85E-3	3.35E-4	1.85E-4
calamari	3.5E-5	1.5E-5	1.5E-5
Priors	0.33	0.33	0.33
$P(\mathcal{W}_d \vert c_i)P(c_i)$	**3.16E-8**	4.35E-14	9.03E-15
$Log_{10}()$	**-10.50**	-13.36	-14.04