NB-NLP
Naive Bayes for Natural Language Processing

Schwartz

April 24, 2017

How do I love thee? Let me count the Bayes

Types of Bayes

- Empirical Bayes
- **Naive Bayes**
- Full Bayes
- Variational Bayes
- Nonparametric Bayes

Types of priors

- Conjugate prior
- Jeffrey's prior
- Improper prior
- (Un)Informative prior
- Objective prior
- Uniform prior

Types of Markov Chain Monte Carlo (MCMC)

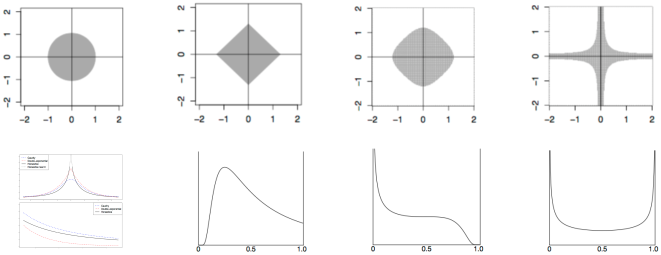
Closed form solutions for posterior distributions are rarely available...

- Gibbs Sampler (cycling through full conditional distributions)
- Metropolis-Hastings (using unnormalized posterior proportionality)
- NUTS: No U-turn sampler (universal probabilistic programming)

Types of Bayesian regularization priors

- Normal-Normal conjugate prior: *ridge regression/regularization*
- Laplace prior: *lasso regularization*
- Cauchy prior: *some other kind of regularization*
- Horseshoe prior: *some other other form of regularization*
  The manuscript presenting the "Horseshoe" prior is entitled "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction"

Gausian Laplace Cauchy Horseshoe



Tails

Implied shrinkage prior profiles from none (0) to total (1) shrinkage

Objectives

1. Understand generative versus predictive modeling

Objectives

1. Understand generative versus predictive modeling
2. Understand "covariance matrix difficulties" when $q > n$

Objectives

1. Understand generative versus predictive modeling
2. Understand "covariance matrix difficulties" when $q > n$
3. Understand how "Naive Bayes" comes to the rescue

Objectives

1. Understand generative versus predictive modeling
2. Understand "covariance matrix difficulties" when $q > n$
3. Understand how "Naive Bayes" comes to the rescue
4. Understand what Naive Bayes classification is & how it works

Objectives

1. Understand generative versus predictive modeling
2. Understand “covariance matrix difficulties” when $q > n$
3. Understand how “Naive Bayes” comes to the rescue
4. Understand what Naive Bayes classification is & how it works
5. Understand how Naive Bayes can be applied to NLP problems

Objectives

1. Understand generative versus predictive modeling
2. Understand "covariance matrix difficulties" when $q > n$
3. Understand how "Naive Bayes" comes to the rescue
4. Understand what Naive Bayes classification is & how it works
5. Understand how Naive Bayes can be applied to NLP problems
6. Know that Naive Bayes is super undemanding computationally

Conditional versus joint models

- Conditional/Predictive/Discriminative
  ("outcome given features")

$$f(Y_i|\boldsymbol{x}_i)$$

Conditional versus joint models

- Conditional/Predictive/Discriminative
  ("outcome given features")

$$f(Y_i|\boldsymbol{x}_i)$$

- Joint $\rightarrow$ Generative ("features given outcome")

$$f(\boldsymbol{X}_i|Y_i) \rightarrow f(Y_i, \boldsymbol{X}_i) \rightarrow f(Y_i|\boldsymbol{X}_i)$$

  So we want to model $\boldsymbol{X}_i|Y_i$... in order to get $Y_i|\boldsymbol{X}_i$

Conditional versus joint models

- Conditional/Predictive/Discriminative
  ("outcome given features")

$$f(Y_i|\boldsymbol{x}_i)$$

- Joint $\rightarrow$ Generative ("features given outcome")

$$f(\boldsymbol{X}_i|Y_i) \rightarrow f(Y_i, \boldsymbol{X}_i) \rightarrow f(Y_i|\boldsymbol{X}_i)$$

  So we want to model $\boldsymbol{X}_i|Y_i$... in order to get $Y_i|\boldsymbol{X}_i$

  For categorical $Y_i \in \{k : k = 1, 2, \cdots K\}$

$$\begin{aligned} f(Y_i, \boldsymbol{X}_i) &= \sum_{k=1}^{K} \Pr(Y_i = k) f(\boldsymbol{X}_i|Y_i = k) \\ &= \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{X}_i) \end{aligned}$$

  so what we need is $\implies \quad f(\boldsymbol{X}_i|Y_i = k) \equiv f_k(\boldsymbol{X}_i)$

Multivariate Normal (MVN) and Multinomial (MN)

$$\left[\begin{array}{cccc}
X_{11} & X_{12} & \cdots & X_{1q} \\
X_{21} & X_{22} & \cdots & X_{2q} \\
\vdots & \vdots & \ddots & \vdots \\
\hline
X_{i1} & X_{i2} & \cdots & X_{iq} \\
\hline
\vdots & \vdots & \ddots & \vdots \\
X_{n1} & X_{n2} & \cdots & X_{nq}
\end{array}\right]$$

$$\boldsymbol{X}_i = (X_{i1}, X_{i2}, \cdots, X_{iq})^T \sim MVN(\boldsymbol{\mu}_{q\times 1}, \Sigma_{q\times q})$$

$$(2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{X}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu})}$$

$$\boldsymbol{X}_i = (X_{i1}, X_{i2}, \cdots, X_{iq})^T \sim MN(\boldsymbol{p}_{q\times 1}, n_i)$$

$$\frac{n_i!}{\prod_{j=1}^{p} X_{ij}!} \prod_{j=1}^{q} p_j^{X_{ij}}$$

Multivariate Models

- $\boldsymbol{X}_i \sim MVN(\boldsymbol{\mu}_{q\times 1}, \Sigma_{q\times q})$
  i.e.,

$$\boldsymbol{X}_i \equiv \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mu_{X_1} \\ \mu_{X_2} \\ \vdots \\ \mu_{X_q} \end{bmatrix}, \begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_q} \\ \sigma_{X_2X_1} & \sigma^2_{X_2} & \cdots & \sigma_{X_2X_q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_qX_1} & \sigma_{X_qX_2} & \cdots & \sigma^2_{X_q} \end{bmatrix} \right)$$

Multivariate Models

- $\boldsymbol{X}_i \sim MVN(\boldsymbol{\mu}_{q\times 1}, \Sigma_{q\times q})$

  i.e.,

$$\boldsymbol{X}_i \equiv \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mu_{X_1} \\ \mu_{X_2} \\ \vdots \\ \mu_{X_q} \end{bmatrix}, \begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_q} \\ \sigma_{X_2X_1} & \sigma^2_{X_2} & \cdots & \sigma_{X_2X_q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_qX_1} & \sigma_{X_qX_2} & \cdots & \sigma^2_{X_q} \end{bmatrix} \right)$$

- What is n? And how do we estimate $\mu_{X_j}, \sigma^2_{X_j}, \sigma_{X_jX_{j'}}$?

Multivariate Models

- $\boldsymbol{X}_i \sim MVN(\boldsymbol{\mu}_{q\times 1}, \Sigma_{q\times q})$
  i.e.,

$$\boldsymbol{X}_i \equiv \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mu_{X_1} \\ \mu_{X_2} \\ \vdots \\ \mu_{X_q} \end{bmatrix}, \begin{bmatrix} \sigma^2_{X_1} & 0 & \cdots & 0 \\ 0 & \sigma^2_{X_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2_{X_q} \end{bmatrix} \right)$$

- What does this matrix specify? And why use it?

Multivariate Models

- $\boldsymbol{X}_i \sim MN(\boldsymbol{p}_{q\times 1}, n_i)$
  i.e.,
  $$\boldsymbol{X}_i \equiv \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \sim MN\left(\begin{bmatrix} p_{X_1} \\ p_{X_2} \\ \vdots \\ p_{X_q} \end{bmatrix}, n_i\right)$$
- Multinomial model counts are (in)dependent?

Multivariate Models

- $\boldsymbol{X}_i \sim MN(\boldsymbol{p}_{q\times 1}, n_i)$

  i.e.,

$$\boldsymbol{X}_i \equiv \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \sim MN\left(\begin{bmatrix} p_{X_1} \\ p_{X_2} \\ \vdots \\ p_{X_q} \end{bmatrix}, n_i \right)$$

- Multinomial model counts are (in)dependent?
- What can we model with this?

Multivariate Models

- $\boldsymbol{X}_i \sim MN(\boldsymbol{p}_{q\times 1}, n_i)$
  i.e.,
  $$\boldsymbol{X}_i \equiv \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \sim MN\left(\begin{bmatrix} p_{X_1} \\ p_{X_2} \\ \vdots \\ p_{X_q} \end{bmatrix}, n_i \right)$$
- Multinomial model counts are (in)dependent?
- What can we model with this? How 'bout text documents?

Multiple Multivariate Models

- $\boldsymbol{X}_i \sim MN(\boldsymbol{p}_{q\times 1}, n_i)$

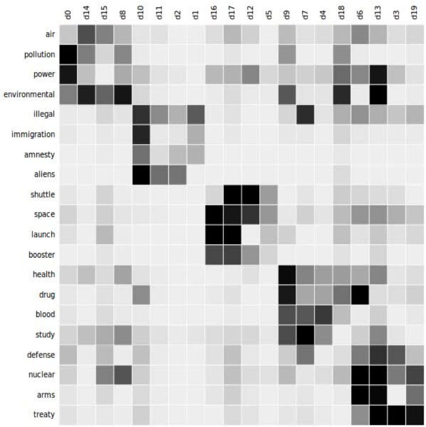
  i.e.,

$$\boldsymbol{X}_i \equiv \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \sim MN\left(\begin{bmatrix} p_{kX_1} \\ p_{kX_2} \\ \vdots \\ p_{kX_q} \end{bmatrix}, n_i \right)$$

- Multinomial model counts are (in)dependent?
- What can we model with this? How 'bout text documents?

Multiple Multivariate Models

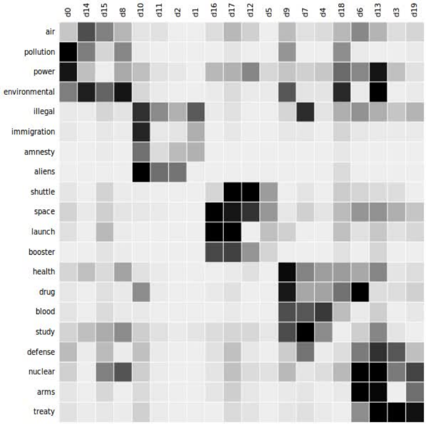
$$\boldsymbol{X}_i \equiv \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \sim MN\left(\begin{bmatrix} p_{kX_1} \\ p_{kX_2} \\ \vdots \\ p_{kX_q} \end{bmatrix}, n_i \right)$$



Multiple Multivariate Models

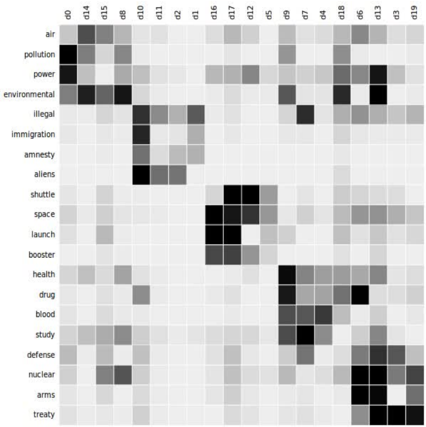
$$\boldsymbol{X}_i \equiv \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \sim MN\left(\begin{bmatrix} p_{kX_1} \\ p_{kX_2} \\ \vdots \\ p_{kX_q} \end{bmatrix}, n_i \right)$$

$$\begin{aligned} & f(Y_i, \boldsymbol{X}_i) \\ = & \sum_{k=1}^{K} \Pr(Y_i = k) f(\boldsymbol{X}_i | Y_i = k) \\ = & \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{X}_i) \end{aligned}$$



Multiple Multivariate Models

$$\boldsymbol{X}_i \equiv \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \sim MN\left(\begin{bmatrix} p_{kX_1} \\ p_{kX_2} \\ \vdots \\ p_{kX_q} \end{bmatrix}, n_i \right)$$



$$f(Y_i, \boldsymbol{X}_i)$$

$$= \sum_{k=1}^{K} \Pr(Y_i = k) f(\boldsymbol{X}_i | Y_i = k)$$

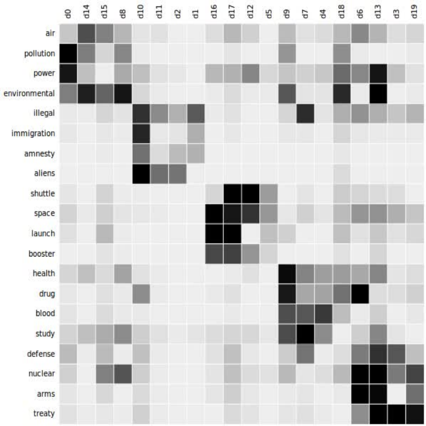
$$= \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{X}_i)$$

$\pi_k \equiv \Pr(Y_i = k)$
estimated with

$$\frac{1}{n} \sum 1_{[Y_i = k]}$$

Multiple Multivariate Models

$$\boldsymbol{X}_i \equiv \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \sim MN\left(\begin{bmatrix} p_{kX_1} \\ p_{kX_2} \\ \vdots \\ p_{kX_q} \end{bmatrix}, n_i \right)$$



$$f(Y_i, \boldsymbol{X}_i)$$

$$= \sum_{k=1}^{K} \Pr(Y_i = k) f(\boldsymbol{X}_i | Y_i = k)$$

$$= \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{X}_i)$$

$\pi_k \equiv \Pr(Y_i = k)$
estimated with

$$\frac{1}{n} \sum 1_{[Y_i = k]}$$

$f_k(\boldsymbol{X}_i) \equiv f(\boldsymbol{X}_i | Y_i = k)$
estimated with

$$MN\left(\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \middle| \begin{bmatrix} \hat{p}_{kX_1} \\ \hat{p}_{kX_2} \\ \vdots \\ \hat{p}_{kX_q} \end{bmatrix}, n_i \right)$$

Why is this even called *Bayes*? Why is this called *Naive*?

$$
\begin{aligned}
& f(Y_i, \boldsymbol{X}_i) \\
&= \sum_{k=1}^{K} \Pr(Y_i = k) f(\boldsymbol{X}_i | Y_i = k) \\
&= \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{X}_i)
\end{aligned}
$$

$\pi_k \equiv \Pr(Y_i = k)$
estimated with

$$\frac{1}{n} \sum 1_{[Y_i = k]}$$

$f_k(\boldsymbol{X}_i) \equiv f(\boldsymbol{X}_i | Y_i = k)$
estimated with

$$MN\left(\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \middle| \begin{bmatrix} \hat{p}_{kX_1} \\ \hat{p}_{kX_2} \\ \vdots \\ \hat{p}_{kX_q} \end{bmatrix}, n_i \right)$$

Why is this even called *Bayes*? Why is this called *Naive*?

$$\Pr(Y_i = k | \boldsymbol{X}_i)$$
$$= \frac{f(\boldsymbol{X}_i | Y_i = k) \Pr(Y = k)}{f(\boldsymbol{X}_i)}$$

$$f(Y_i, \boldsymbol{X}_i)$$
$$= \sum_{k=1}^{K} \Pr(Y_i = k) f(\boldsymbol{X}_i | Y_i = k)$$
$$= \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{X}_i)$$

$\pi_k \equiv \Pr(Y_i = k)$
estimated with

$$\frac{1}{n} \sum 1_{[Y_i = k]}$$

$f_k(\boldsymbol{X}_i) \equiv f(\boldsymbol{X}_i | Y_i = k)$
estimated with

$$MN\left(\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \middle| \begin{bmatrix} \hat{p}_{kX_1} \\ \hat{p}_{kX_2} \\ \vdots \\ \hat{p}_{kX_q} \end{bmatrix}, n_i \right)$$

Why is this even called *Bayes*? Why is this called *Naive*?

$$
\begin{aligned}
&\Pr(Y_i = k|\boldsymbol{X}_i)\\
=&\frac{f(\boldsymbol{X}_i|Y_i = k)\Pr(Y = k)}{f(\boldsymbol{X}_i)}\\
=&\frac{\pi_k f_k(\boldsymbol{X}_i)}{\sum_{k'=1}^{K}\pi_{k'} f_{k'}(\boldsymbol{X}_i)}
\end{aligned}
$$

$$
\begin{aligned}
&f(Y_i, \boldsymbol{X}_i)\\
=&\sum_{k=1}^{K}\Pr(Y_i = k) f(\boldsymbol{X}_i|Y_i = k)\\
=&\sum_{k=1}^{K}\pi_k f_k(\boldsymbol{X}_i)
\end{aligned}
$$

$\pi_k \equiv \Pr(Y_i = k)$
estimated with

$$\frac{1}{n}\sum 1_{[Y_i = k]}$$

$f_k(\boldsymbol{X}_i) \equiv f(\boldsymbol{X}_i|Y_i = k)$
estimated with

$$MN\left(\begin{bmatrix} X_1\\ X_2\\ \vdots\\ X_q\end{bmatrix} \middle| \begin{bmatrix} \hat{p}_{kX_1}\\ \hat{p}_{kX_2}\\ \vdots\\ \hat{p}_{kX_q}\end{bmatrix}, n_i\right)$$

Why is this even called *Bayes*? Why is this called *Naive*?

$$\Pr(Y_i = k|\boldsymbol{X}_i)$$

$$= \frac{f(\boldsymbol{X}_i|Y_i = k)\Pr(Y = k)}{f(\boldsymbol{X}_i)}$$

$$= \frac{\pi_k f_k(\boldsymbol{X}_i)}{\sum_{k'=1}^{K} \pi_{k'} f_{k'}(\boldsymbol{X}_i)}$$

$$\propto \pi_k f_k(\boldsymbol{X}_i)$$

$$f(Y_i, \boldsymbol{X}_i)$$

$$= \sum_{k=1}^{K} \Pr(Y_i = k) f(\boldsymbol{X}_i|Y_i = k)$$

$$= \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{X}_i)$$

$\pi_k \equiv \Pr(Y_i = k)$
estimated with

$$\frac{1}{n}\sum 1_{[Y_i=k]}$$

$f_k(\boldsymbol{X}_i) \equiv f(\boldsymbol{X}_i|Y_i = k)$
estimated with

$$MN\left(\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \middle| \begin{bmatrix} \hat{p}_{kX_1} \\ \hat{p}_{kX_2} \\ \vdots \\ \hat{p}_{kX_q} \end{bmatrix}, n_i\right)$$

Why is this even called *Bayes*? Why is this called *Naive*?

$$\Pr(Y_i = k|\boldsymbol{X}_i)$$
$$= \frac{f(\boldsymbol{X}_i|Y_i = k)\Pr(Y = k)}{f(\boldsymbol{X}_i)}$$

$$= \frac{\pi_k f_k(\boldsymbol{X}_i)}{\sum_{k'=1}^{K} \pi_{k'} f_{k'}(\boldsymbol{X}_i)}$$
$$\propto \pi_k f_k(\boldsymbol{X}_i)$$
$$\propto \pi_k \hat{p}_{kX_1}^{X_{i1}} \cdot \hat{p}_{kX_2}^{X_{i2}} \cdots \hat{p}_{kX_q}^{X_{ip}}$$

$$f(Y_i, \boldsymbol{X}_i)$$
$$= \sum_{k=1}^{K} \Pr(Y_i = k) f(\boldsymbol{X}_i|Y_i = k)$$
$$= \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{X}_i)$$

$\pi_k \equiv \Pr(Y_i = k)$
estimated with

$$\frac{1}{n}\sum 1_{[Y_i=k]}$$

$f_k(\boldsymbol{X}_i) \equiv f(\boldsymbol{X}_i|Y_i = k)$
estimated with

$$MN\left(\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \middle| \begin{bmatrix} \hat{p}_{kX_1} \\ \hat{p}_{kX_2} \\ \vdots \\ \hat{p}_{kX_q} \end{bmatrix}, n_i\right)$$

Why is this even called *Bayes*? Why is this called *Naive*?

$$
\begin{aligned}
&\Pr(Y_i = k|\boldsymbol{X}_i)\\
&= \frac{f(\boldsymbol{X}_i|Y_i = k)\Pr(Y = k)}{f(\boldsymbol{X}_i)}\\
&= \frac{\pi_k f_k(\boldsymbol{X}_i)}{\sum_{k'=1}^{K} \pi_{k'} f_{k'}(\boldsymbol{X}_i)}\\
&\propto \pi_k f_k(\boldsymbol{X}_i)\\
&\propto \pi_k \hat{p}_{kX_1}^{X_{i1}} \cdot \hat{p}_{kX_2}^{X_{i2}} \cdots \hat{p}_{kX_q}^{X_{ip}}\\
&= \pi_k \prod_{\underline{word} \in words} \hat{p}_{k\underline{word}}
\end{aligned}
$$

$$
\begin{aligned}
&f(Y_i, \boldsymbol{X}_i)\\
&= \sum_{k=1}^{K} \Pr(Y_i = k) f(\boldsymbol{X}_i|Y_i = k)\\
&= \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{X}_i)
\end{aligned}
$$

$\pi_k \equiv \Pr(Y_i = k)$
estimated with

$$\frac{1}{n}\sum 1_{[Y_i = k]}$$

$f_k(\boldsymbol{X}_i) \equiv f(\boldsymbol{X}_i|Y_i = k)$
estimated with

$$MN\left(\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \middle| \begin{bmatrix} \hat{p}_{kX_1} \\ \hat{p}_{kX_2} \\ \vdots \\ \hat{p}_{kX_q} \end{bmatrix}, n_i \right)$$

Why is this even called *Bayes*? Why is this called *Naive*?

$$\Pr(Y_i = k|\boldsymbol{X}_i)$$
$$= \frac{f(\boldsymbol{X}_i|Y_i = k)\Pr(Y = k)}{f(\boldsymbol{X}_i)}$$

$$= \frac{\pi_k f_k(\boldsymbol{X}_i)}{\sum_{k'=1}^{K} \pi_{k'} f_{k'}(\boldsymbol{X}_i)}$$
$$\propto \pi_k f_k(\boldsymbol{X}_i)$$
$$\propto \pi_k \hat{p}_{kX_1}^{X_{i1}} \cdot \hat{p}_{kX_2}^{X_{i2}} \cdots \hat{p}_{kX_q}^{X_{ip}}$$
$$= \pi_k \prod_{\underline{word} \in words} \hat{p}_{k\underline{word}}$$

Notice how probability of a word is independent of previous words

$$f(Y_i, \boldsymbol{X}_i)$$
$$= \sum_{k=1}^{K} \Pr(Y_i = k) f(\boldsymbol{X}_i|Y_i = k)$$
$$= \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{X}_i)$$

$\pi_k \equiv \Pr(Y_i = k)$
estimated with
$$\frac{1}{n}\sum 1_{[Y_i=k]}$$

$f_k(\boldsymbol{X}_i) \equiv f(\boldsymbol{X}_i|Y_i = k)$
estimated with
$$MN\left(\left[\begin{array}{c} X_1 \\ X_2 \\ \vdots \\ X_q \end{array}\right] \middle| \left[\begin{array}{c} \hat{p}_{kX_1} \\ \hat{p}_{kX_2} \\ \vdots \\ \hat{p}_{kX_q} \end{array}\right], n_i\right)$$

Why is this even called *Bayes*? Why is this called *Naive*?

$$
\begin{aligned}
&\Pr(Y_i = k|\boldsymbol{X}_i)\\
&= \frac{f(\boldsymbol{X}_i|Y_i = k)\Pr(Y = k)}{f(\boldsymbol{X}_i)}\\
&= \frac{\pi_k f_k(\boldsymbol{X}_i)}{\sum_{k'=1}^{K} \pi_{k'} f_{k'}(\boldsymbol{X}_i)}\\
&\propto \pi_k f_k(\boldsymbol{X}_i)\\
&\propto \pi_k \hat{p}_{kX_1}^{X_{i1}} \cdot \hat{p}_{kX_2}^{X_{i2}} \cdots \hat{p}_{kX_q}^{X_{ip}}\\
&= \pi_k \prod_{\underline{word} \in words} \hat{p}_{k\underline{word}}
\end{aligned}
$$

Notice how probability of a word

is independent of previous words

$$
\neq \pi_k \prod_{\underline{word} \in words} \hat{p}_{k\underline{word}|\text{previous words}}
$$

$$
\begin{aligned}
&f(Y_i, \boldsymbol{X}_i)\\
&= \sum_{k=1}^{K} \Pr(Y_i = k) f(\boldsymbol{X}_i|Y_i = k)\\
&= \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{X}_i)
\end{aligned}
$$

$\pi_k \equiv \Pr(Y_i = k)$
estimated with

$$
\frac{1}{n} \sum 1_{[Y_i = k]}
$$

$f_k(\boldsymbol{X}_i) \equiv f(\boldsymbol{X}_i|Y_i = k)$
estimated with

$$
MN\left(\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \middle| \begin{bmatrix} \hat{p}_{kX_1} \\ \hat{p}_{kX_2} \\ \vdots \\ \hat{p}_{kX_q} \end{bmatrix}, n_i \right)
$$

Tricks

Laplace Smoothing:

$$\hat{p}_{kX_j} = \frac{\#(\text{times } X_j \text{ appears in Class } k) + \alpha}{\#(\text{words in Class } k) + \alpha \times |\text{Vocab}|}$$

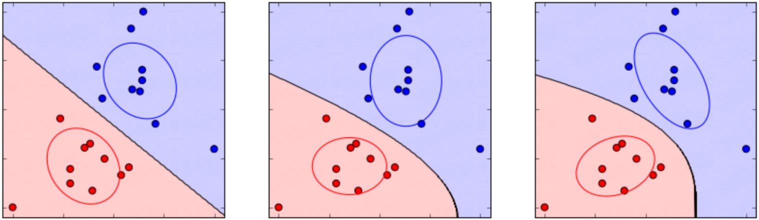
Exponentiating the sum of log probabilities:

$$\begin{aligned}
&\Pr(Y_i = k|\boldsymbol{X}_i) \\
&= \frac{\pi_k \hat{p}_{kX_1}^{X_{i1}} \cdot \hat{p}_{kX_2}^{X_{i2}} \cdots \hat{p}_{kX_q}^{X_{ip}}}{\sum_{k'=1}^{K} \pi_{k'} \hat{p}_{k'X_1}^{X_{i1}} \cdot \hat{p}_{k'X_2}^{X_{i2}} \cdots \hat{p}_{k'X_q}^{X_{ip}}} = \frac{1}{\sum_{k'=1}^{K} \frac{\pi_{k'} \hat{p}_{k'X_1}^{X_{i1}} \cdot \hat{p}_{k'X_2}^{X_{i2}} \cdots \hat{p}_{k'X_q}^{X_{ip}}}{\pi_k \hat{p}_{kX_1}^{X_{i1}} \cdot \hat{p}_{kX_2}^{X_{i2}} \cdots \hat{p}_{kX_q}^{X_{ip}}}} \\
&= \frac{1}{\sum_{k'=1}^{K} \frac{\pi_{k'}}{\pi_k} exp\left(\sum_{j=1}^{q} log(\hat{p}_{k'X_j}^{X_{ij}}) - \sum_{j=1}^{q} log(\hat{p}_{kX_j}^{X_{ij}})\right)}
\end{aligned}$$

What is the assumption on the covariance matrix doing?

Back to the MVN...

$$
\begin{aligned}
&\Pr(Y_i = k|\boldsymbol{X}_i) \\
&\propto \pi_k f_k(\boldsymbol{X}_i) \\
&= \pi_k \prod_{j=1}^{q} f_k(X_{ji})
\end{aligned}
\qquad
\begin{bmatrix}
\hat{\sigma}^2_{kX_1} & 0 & \cdots & 0 \\
0 & \hat{\sigma}^2_{kX_2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \hat{\sigma}^2_{kX_q}
\end{bmatrix}
$$

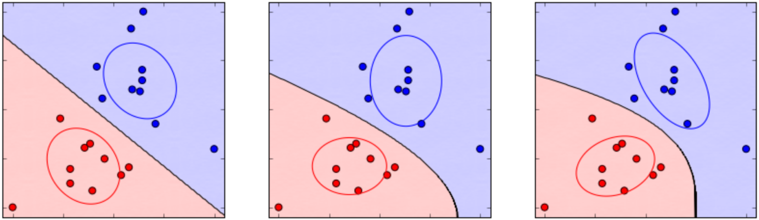


What is the assumption on the covariance matrix doing?

Back to the MVN...

$$\begin{aligned} &\Pr(Y_i = k|\boldsymbol{X}_i) \\ &\propto \pi_k f_k(\boldsymbol{X}_i) \\ &= \pi_k \prod_{j=1}^{q} f_k(X_{ji}) \end{aligned}$$

$$\begin{bmatrix} \hat{\sigma}^2_{kX_1} & 0 & \cdots & 0 \\ 0 & \hat{\sigma}^2_{kX_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\sigma}^2_{kX_q} \end{bmatrix}$$



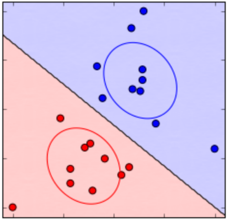
Linear Discriminant Analysis (LDA) Naive Bayes Quadratic Discriminant Analysis (QDA)

What is the assumption on the covariance matrix doing?

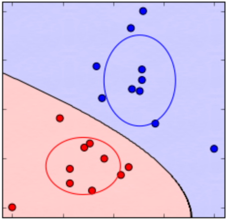
Back to the MVN...

$$\begin{aligned} &\Pr(Y_i = k|\boldsymbol{X}_i) \\ &\propto \pi_k f_k(\boldsymbol{X}_i) \\ &= \pi_k \prod_{j=1}^{q} f_k(X_{ji}) \end{aligned}$$

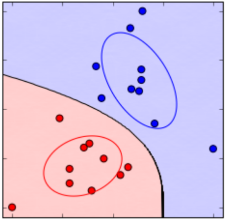
$$\begin{bmatrix} \hat{\sigma}^2_{kX_1} & 0 & \cdots & 0 \\ 0 & \hat{\sigma}^2_{kX_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\sigma}^2_{kX_q} \end{bmatrix}$$



Linear Discriminant Analysis (LDA)



Naive Bayes



Quadratic Discriminant Analysis (QDA)

$\Uparrow\Uparrow\Uparrow\Uparrow\Uparrow$

Parting comments

- Truly correlated features can hamper Naive Bayes (NB)

Parting comments

- Truly correlated features can hamper Naive Bayes (NB)
- Probability estimates are unreliable with the naive assumption

Parting comments

- Truly correlated features can hamper Naive Bayes (NB)
- Probability estimates are unreliable with the naive assumption
- But NB classifications can be workable... but

Parting comments

- Truly correlated features can hamper Naive Bayes (NB)
- Probability estimates are unreliable with the naive assumption
- But NB classifications can be workable... but
- NB is typically outperformed by less naive methodologies

Parting comments

- Truly correlated features can hamper Naive Bayes (NB)
- Probability estimates are unreliable with the naive assumption
- But NB classifications can be workable... but
- NB is typically outperformed by less naive methodologies

  However...

Parting comments

- Truly correlated features can hamper Naive Bayes (NB)
- Probability estimates are unreliable with the naive assumption
- But NB classifications can be workable... but
- NB is typically outperformed by less naive methodologies

  However...
- NB is super undemanding computationally:

Parting comments

- Truly correlated features can hamper Naive Bayes (NB)
- Probability estimates are unreliable with the naive assumption
- But NB classifications can be workable... but
- NB is typically outperformed by less naive methodologies

  However...
- NB is super undemanding computationally:

  NB can handle huge data sets very quickly – i.e., in real time

Parting comments

- Truly correlated features can hamper Naive Bayes (NB)
- Probability estimates are unreliable with the naive assumption
- But NB classifications can be workable... but
- NB is typically outperformed by less naive methodologies

  However...
- NB is super undemanding computationally:

  NB can handle huge data sets very quickly – i.e., in real time

  NB can handle wide data sets other methodologies can't...

Parting comments

- Truly correlated features can hamper Naive Bayes (NB)
- Probability estimates are unreliable with the naive assumption
- But NB classifications can be workable... but
- NB is typically outperformed by less naive methodologies

  However...
- NB is super undemanding computationally:

  NB can handle huge data sets very quickly – i.e., in real time

  NB can handle wide data sets other methodologies can't...
- And NB is very simple to implement and use...

Parting comments

- Truly correlated features can hamper Naive Bayes (NB)
- Probability estimates are unreliable with the naive assumption
- But NB classifications can be workable... but
- NB is typically outperformed by less naive methodologies

  However...
- NB is super undemanding computationally:

  NB can handle huge data sets very quickly – i.e., in real time

  NB can handle wide data sets other methodologies can't...
- And NB is very simple to implement and use...

  Although isn't *everything* in scikit-learn?