Bayesian Classification Methods

Suchit Mehrotra

North Carolina State University

smehrot@ncsu.edu

October 24, 2014

How do you define probability?

There exists a fundamental difference between how Frequentists and Bayesians view probability:

- **Frequentist:** Think of probability as choosing a random point in a Venn diagram and determing the percentage of times the point would fall in a certain set as you repeat this process an infinite number of times. Seeks to be "objective".
- **Bayesian:** The probability assigned to an event depends person to person. It is the condition under which a person would make a bet with their own money. Reflects a state of belief.

Conditional Probability and Bayes Rule

The fundamental mechanism through which a Bayesian would update their beliefs about a certain event is Bayes Rule.

Let A and B be events.

- Conditional Probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ where } P(B) > 0 \implies$$

$$P(A \cap B) = P(B|A)P(A)$$

- Bayes Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Posterior Distributions

Let θ be a parameter of interest and $\mathbf{y}$ be observed data. The Frequentist approach considers θ to be fixed and the data to be random, whereas Bayesians view θ as a random variable and $\mathbf{y}$ as fixed.

- The posterior distribution for θ is an application of Bayes Rule:

$$\pi(\theta|\mathbf{y}) = \frac{f_{\mathbf{Y}|\theta}(\mathbf{y}|\theta)f_{\Theta}(\theta)}{f_{\mathbf{Y}}(y)}$$

- $f_{\Theta}(\theta)$ encapsulates our prior belief about the parameter. Our prior beliefs are then updated by the data we observe.
- Since θ is treated as a random variable, we can now make probability statements about it.

Primary Goals

Frequentist:

- Calculate point estimates for the parameters and use standard errors to calculate confidence intervals.

Bayesian:

- Derive the posterior distribution of the parameter.
- Summarize this distribution with information about mean, median, mode, quantile, variance etc.
- Used credible intervals to show ranges which contain the parameter with high probability.

Bayesian Learning

- The Bayesian framework provides us with a mathematically rigorous way to update our beliefs as we gather new information.
- There are no restrictions on what can constitute prior information. Therefore we can use posterior distributions from time 1 as our prior for time 2.

$$\begin{aligned}
\pi_1(\theta|\mathbf{y}_1) &\propto f_{\mathbf{Y}|\theta}(\mathbf{y}_1|\theta)\ f_{\Theta}(\theta) \\
\pi_2(\theta|\mathbf{y}_1, \mathbf{y}_2) &\propto f_{\mathbf{Y}|\theta}(\mathbf{y}_2|\theta)\ \pi_1(\theta|\mathbf{y}_1) \\
&\vdots \\
\pi_n(\theta|\mathbf{y}_1, \ldots, \mathbf{y}_n) &\propto f_{\mathbf{Y}|\theta}(\mathbf{y}_n|\theta)\ \pi_{n-1}(\theta|\mathbf{y}_1, \ldots, \mathbf{y}_{n-1})
\end{aligned}$$

Overview of Differences [Gill (2008)]

	Interpretation of Probability
Frequentist:	Observed result from infinite series of trials performed or imagined under identical conditions.
	Probabilistic quantity of interest is $p(data\|H_o)$
Bayesian:	Probability is the researcher/observer ”degree of belief” before or after the data are observed.
	Probabilistic quantity of interest is $p(\theta\|data)$.
	What is Fixed and Variable
Frequentist:	Data are a iid random sample from continuous stream.
	Parameters are fixed by nature.
Bayesian:	Data observed and so fixed by the sample generated.
	Parameters are unknown and described distributionally.
	How are Results Summarized?
Frequentist:	Point estimates and standard errors.
	95% confidence intervals indicating that 19/20 times the interval covers the true parameter value, on average.
Bayesian:	Descriptions of posteriors such as means and quantiles.
	Highest posterior density intervals indicating region of highest probability.

Linear Regression (Frequentist)

- Consider the linear model $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ where $\mathbf{X}$ is a n x k matrix with rank k, β is a k x 1 vector of coefficients and $\mathbf{y}$ is an n x 1 vector of responses. $\mathbf{e}$ is a vector of errors $\sim$ iid N(0, $\sigma^2\mathbf{I}$).
- Frequentist regression seeks point estimates by maximizing likelihood function with respect to β and σ^2.

$$L(\beta, \sigma^2 | \mathbf{X}, \mathbf{y}) = (2\pi\sigma^2)^{\frac{-n}{2}} exp\left[\frac{-1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\right]$$

- The unbiased OLS estimates are:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \qquad \hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})}{n-k}$$

Bayesian Regression

Bayesian regression, on the other hand, finds the posterior distribution of the parameters β and σ^2.

$$\begin{aligned}\pi(\boldsymbol{\beta}, \sigma^2 \,|\, \mathbf{X}, \mathbf{y}) &\propto L(\boldsymbol{\beta}, \sigma^2 \,|\, \mathbf{X}, \mathbf{y}) f_{\boldsymbol{\beta}, \sigma^2}(\boldsymbol{\beta},\, \sigma^2) \\ &\propto L(\boldsymbol{\beta}, \sigma^2 \,|\, \mathbf{X}, \mathbf{y}) f_{\boldsymbol{\beta}|\sigma^2}(\boldsymbol{\beta} \,|\, \sigma^2) f_{\sigma^2}(\sigma^2)\end{aligned}$$

Setup	Prior	Posterior
Conjugate	$\boldsymbol{\beta} \,\vert\, \sigma^2 \sim N(\mathbf{b}, \sigma^2)$ $\sigma^2 \sim \mathcal{IG}(a, b)$	$\boldsymbol{\beta} \,\vert\, \mathbf{X} \sim t(n+a-k)$ $\sigma^2 \,\vert\mathbf{X} \sim \mathcal{IG}(n+a-k, \frac{1}{2}\hat{\sigma}^2(n+a-k))$
Uninformative	$\boldsymbol{\beta} \propto c$ over $[-\infty, \infty]$ $\sigma^2 = \frac{1}{\sigma}$ over $[0:\infty]$	$\boldsymbol{\beta} \,\vert\, \mathbf{X} \sim t(n-k)$ $\sigma^2 \,\vert\, \mathbf{X} \sim \mathcal{IG}(\frac{1}{2}(n-k-1), \frac{1}{2}\hat{\sigma}^2(n-k))$

Gill(2008)

Additionally, the posterior distribution conditioned on σ^2 for β in the non-informative case:

$$\boldsymbol{\beta}|\sigma^2, \mathbf{y} \sim N(\hat{\boldsymbol{\beta}}, \sigma^2 V_{\boldsymbol{\beta}}) \quad \text{where} \quad V_{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}$$

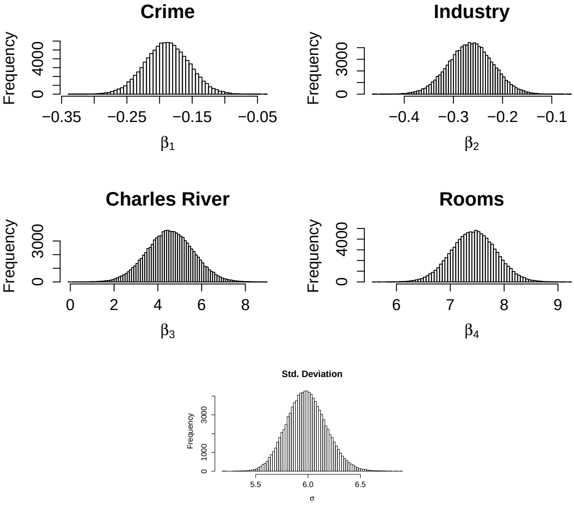
Example: Median Home Prices

- From the MASS library: Data regarding 506 housing values in suburbs of Boston
- Fit a model with *medv* (median value of owner occupied homes) as the response and four predictors.
- Implementation with the *blinreg* function in the LearnBayes package.

```
boston.fit <- lm(medv ~ crim + indus + chas + rm, data = Boston, y = TRUE,
    x = TRUE)

theta.sample <- blinreg(boston.fit$y, boston.fit$x, 100000)
```

Example: Posterior Samples of Parameters



Example: Confidence Intervals for Parameters

Bayesian regression with non-informative priors leads to similar estimates as OLS.

- **OLS Confidence Intervals:**

	2.5 %	97.5 %
(Intercept)	-26.43	-15.29
crim	-0.26	-0.12
indus	-0.35	-0.18
chas	2.48	6.65
rm	6.62	8.25

- **Bayesian Credible Intervals:**

	Intercept	crim	indus	chas	rm
2.5%	-26.42	-0.26	-0.35	2.48	6.62
97.5%	-15.30	-0.12	-0.18	6.65	8.25

Logistic Regression

Like Linear Regression, Logistic Regression is also a likelihood maximization problem in the frequentist setup. When the number of parameters is two, the log-likelihood function is:

$$\ell(\beta_0, \beta_1 | y) = \beta_0 \sum_{i=1}^{n} y_i + \beta_1 \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} log(1 + e^{\beta_0 + \beta_1 x_i})$$

In the Bayesian setting, we incorporate prior information and find the posterior distribution of the parameters:

$$\pi(\boldsymbol{\beta} | \mathbf{X}, \mathbf{y}) \propto L(\boldsymbol{\beta} | \mathbf{X}, \mathbf{y}) f_{\boldsymbol{\beta}}(\boldsymbol{\beta})$$

The standard "weakly-informative" prior used in the *arm* package is the Student-t distribution with 1 degree of freedom (Cauchy distribution).

Logistic Regression: Example with Cancer Data Set

Goal: Run Logistic Regression on the cancer data set from the University of Wisconsin using the *bayesglm* function in the *arm* package.

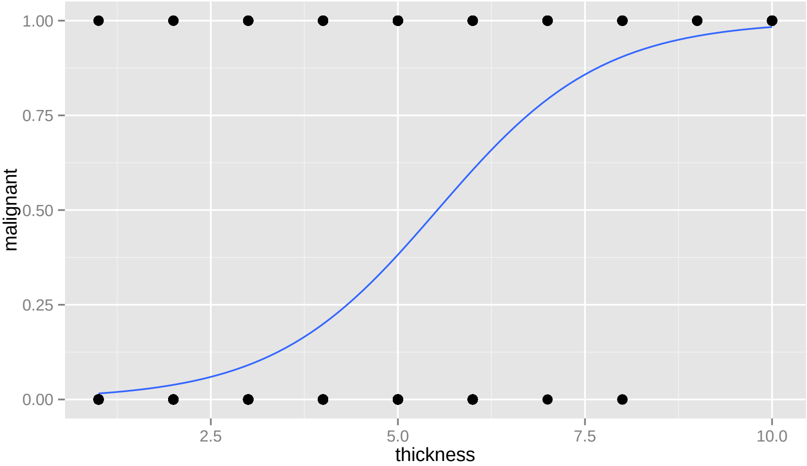
```
cancer.bayesglm <- bayesglm(malignant ˜ . , data = cancer, family = binomial
    (link = "logit"))
```

Using this function, you can get point estimates (posterior modes) for your parameters and their standard errors. These estimates will be the same as the output given by *glm* in most cases without substantive prior information.

	Estimate	Std. Error
(Intercept)	-8.2409	1.2842
id	0.0000	0.0000
thickness	0.4522	0.1270
unif.size	0.0063	0.1676
unif.shape	0.3642	0.1876
adhesion	0.2002	0.1067
cell.size	0.1460	0.1458
bare.nuclei1	-1.1499	0.8526

Some estimates from *bayesglm*

Logistic Regression: Example with Cancer Data Set



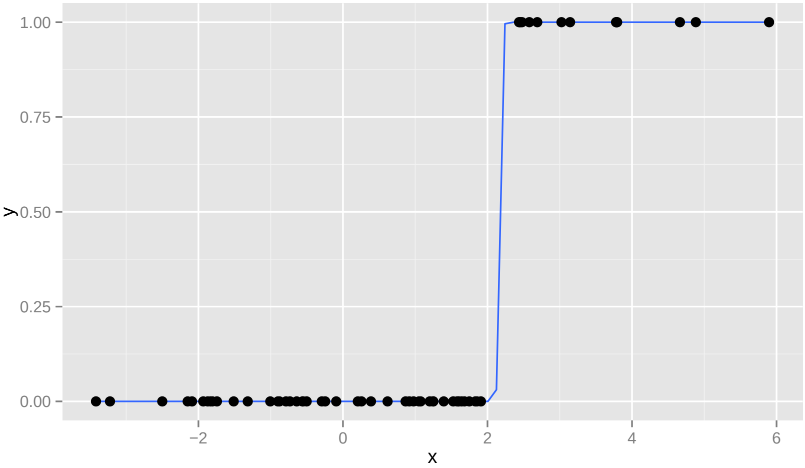
Logistic Regression: Separation

Separation occurs when a particular linear combination of predictors is always associated with a particular response.

- Can lead to maximum likelihood estimates of ∞ for some parameters!
- A common fix is to drop predictors from the model.
  - Can lead to the best predictors being dropped from the model [Zorn(2005)]
- 2008 paper by Gelman, Jakulin, Pittau & Su proposes a Bayesian fix.
  - Use the t-distrubtion as the "weakly-informative" prior.
  - Created *bayesglm* in the *arm* (Applied Regression and Multilevel Modeling) package with the t-distribution as the default prior.

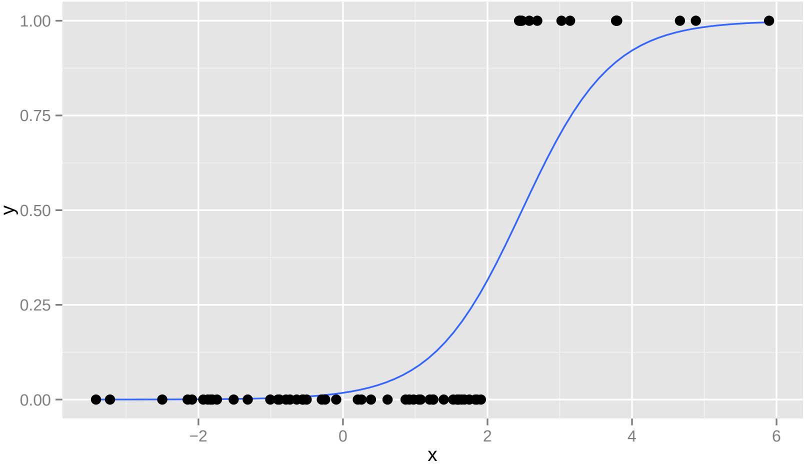
Using *bayesglm* instead of *glm* can solve this problem!

Logistic Regression: Separation Example



Logistic Regression: Separation Example

This problem is fixed using *bayesglm*



Logistic Regression: Separation Example

	Estimate	Std. Error	z value	$\Pr(>\|z\|)$
(Intercept)	-164.0925	69589.0320	-0.00	0.9981
x	75.6372	32263.2265	0.00	0.9981

glm fit

	Estimate	Std. Error
(Intercept)	-4.0144	0.9054
x	1.6192	0.3867

bayesglm fit

Naive Bayes: Overview

Like Linear Discriminant Analysis, Naive Bayes uses Bayes Rule to determine the probability that a particular outcome is in class C_ℓ.

$$P(Y = C_\ell | X) = \frac{P(X|Y = C_\ell)\, P(Y = C_\ell)}{P(X)}$$

- $P(Y = C_\ell | X)$ is the posterior probability of the class given a set of predictors
- $P(Y = C_\ell)$ is the prior probability of the outcome
- $P(X)$ is the probability of the predictors
- $P(X|Y = C_\ell)$ is the probablity of observing a set of predictors given that Y belongs to class ℓ.

Naive Bayes: Independence Assumption

$P(\mathbf{X}|Y = C_\ell)$ is extremely difficult to calculate without a strong set of assumptions. The Naive Bayes model simplifies this calculation with the assumption that all predictors are mutually independent given Y.

$$\begin{aligned}
P(Y = C_\ell \,|\, X_1 \ldots X_n) &= \frac{P(Y = C_\ell)\, P(X_1, \ldots, X_n \,|\, Y = C_\ell)}{P(X_1, \ldots, X_n)} \\
&\propto P(Y = C_\ell)\, P(X_1, \ldots, X_n \,|\, Y = C_\ell) \\
&\propto P(Y = C_\ell)\, P(X_1 \,|\, Y = C_\ell) \ldots P(X_n \,|\, Y = C_\ell) \\
&\propto P(Y = C_\ell) \prod_{i=1}^{n} P(X_i \,|\, Y = C_\ell)
\end{aligned}$$

Naive Bayes: Class Probabilities and Assignment

- The classifier assigns the class which has the highest posterior probability:

$$\underset{C_\ell}{\operatorname{argmax}}\ P(Y = C_\ell) \prod_{i=1}^{n} P(X_i = x_i \mid Y = C_\ell)$$

- We need to assume prior probabilites for class priors. We have two basic options.
  - Equiprobable probabilites: $P(C_\ell) = \frac{1}{n}$
  - Calculation from training set: $P(C_\ell) = \frac{\text{number of samples in class } \ell}{\text{total number of samples}}$
- The third option is to use information based on prior beliefs about the probablity that $Y = C_\ell$

Naive Bayes: Distributional Assumptions

Even though the independence assumptions simplifies the classifier, we still need to make assumptions about the predictors.

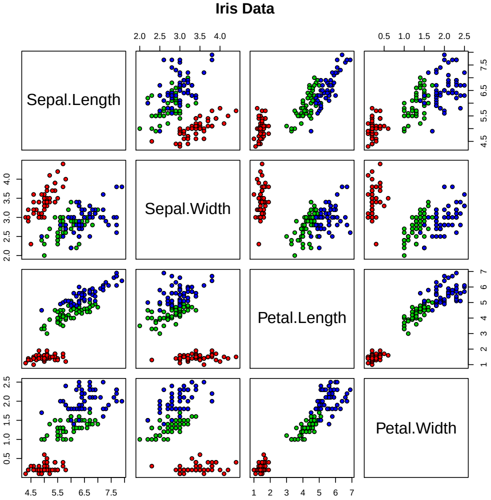
- In the case of continuous predictors, the class probabilities are calculated using a Normal Distribution by default.
  - The classifier estimates the μ_ℓ and σ_ℓ for each predictor,

$$P(X = x|Y = C_\ell) = \frac{1}{\sigma_\ell\sqrt{2\pi}} exp\left[\frac{-1}{2\sigma_\ell^2}(x - \mu_\ell)^2\right]$$

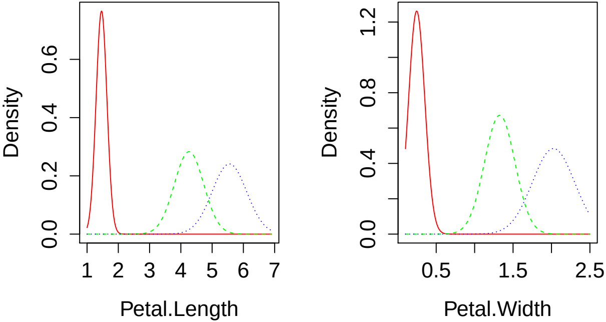
- Other choices include:
  - Multinomial
  - Bernoulli
  - Semi-supervised methods

Naive Bayes: Simple Example

Classify the Iris dataset using the NaiveBayes function in the klaR package.



Naive Bayes: Simple Example



	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	47	3
virginica	0	3	47

Naive Bayes: Non-Parametric Estimation

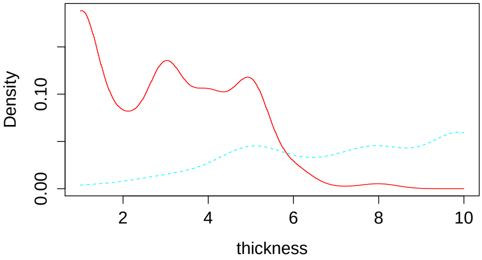
```
cancer.nb <- NaiveBayes(malignant ~. ,data = cancer, usekernel = TRUE )
```

	0	1
0	445	9
1	12	232

TRUE (97%)

	0	1
0	434	6
1	23	235

FALSE (96%)



Naive Bayes: Priors

Bayesian approaches allow for the inclusion of prior information into the model. In this context priors are applied to the probability of belonging to a group.

	0	1
0	442	3
1	15	238

P(0) = 0.1; P(1) = 0.9

	0	1
0	443	4
1	14	237

P(0) = 0.25; P(1) = 0.75

	0	1
0	444	7
1	13	234

P(0) = 0.5; P(1) = 0.5

	0	1
0	445	18
1	12	223

P(0) = 0.9; P(1) = 0.1

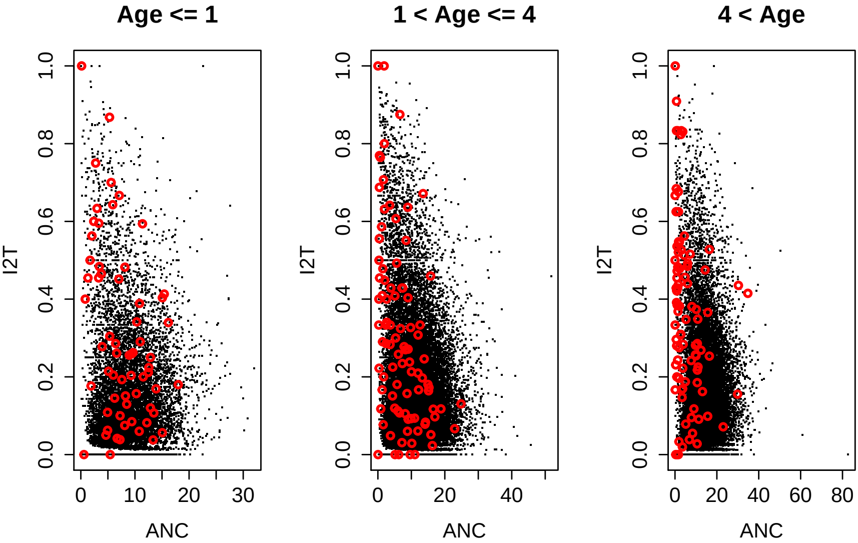
Naive Bayes: Sepsis Example

Fit Naive Bayes model on data regarding whether or not a child develops sepsis based on three predictors. Data from research done at Kaiser Permanente (Courtesy of Dr. David Draper at UCSC).

- Response is whether or not a child developed Sepsis (66,846 observations).
    - This is an extremely bad outcome for a child. They either die or go into a permanent coma.
- Two predictors ANC and I2T are derived from a blood test called the Complete Blood Count.
- Third predictor is the baby's age in hours at which the blood test results were obtained.

In this scenario, our goal is likely to predict accurately which children will get sepsis. Overall classification rate is likely not as important.

Naive Bayes: Sepsis Example



Naive Bayes: Sepsis Example

As can be seen by the confusion matrices below, the overall classification rate of around 98.6% does not make up for the fact that we only correctly classify the sepsis incidence 11% of the time.

	0	1
0	65887	216
1	715	28

usekernel = FALSE

	0	1
0	66589	239
1	13	5

usekernel = TRUE

Naive Bayes: Pros and Cons

- **Advantages**
  - Due to the independence assumption, there is no curse of dimensionality! We can fit models with p $>$ n without any issues.
  - Independence assumption simplifies math and speeds up the model fitting process.
  - We can include prior information to increase accuracy.
  - Have the option of not making distributional assumptions about each predictor.
- **Disadvantages**
  - If independence assumptions do not hold, this model can perform poorly.
  - Non-parametric approach can overfit data.
  - Our prior information could be wrong.

Logistic Regression with Sepsis Data

How does Logistic regression fare with the Sepsis data?

	0	1
0	66510	233
1	92	11

Cutoff: 0.1, % Correct: 99.51%

	0	1
0	66245	217
1	357	27

Cutoff: 0.05, % Correct: 99.14%

	0	1
0	65582	180
1	1020	64

Cutoff: 0.025, % Correct: 98.20%

	0	1
0	63339	135
1	3263	109

Cutoff: 0.01, % Correct: 94.91%

The End

Questions?