CS 335: Bayesian Reasoning

Dan Sheldon

November 20, 2014

Bayes Rule

Let A and B be two events. Then:

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

(Derivation: apply definition of conditional probability twice)

Interpretation I

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

$A =$ hypothesis

$B =$ evidence

$P(A)$: **prior probability** of hypothesis

$P(A|B)$: **posterior probability** of hypothesis given evidence

$P(B|A)$: **likelihood** of evidence given hypothesis

Example:

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

$A =$ "has cancer"

$B =$ "smokes"

What is $P(\text{has cancer}|\text{smokes})$?

Can obtain from:

- $P(\text{smokes})$, $P(\text{has cancer})$ (population stats)
- $P(\text{smokes}|\text{has cancer})$ (stats from cancer patients)

Bayes Rule II

Suppose $A_1, \ldots, A_k$ are competing hypotheses (events that partition Ω)

$$\boxed{P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(B)}}$$

Apply law of total probability to denominator to get a more useful form:

$$\boxed{P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + \ldots + P(A_k)P(B|A_k)}}$$

Interpretation II

$$\boxed{P(A_i|B) = \frac{P(A_i)P(B|A_i)}{P(A_1)P(B|A_1) + \ldots + P(A_k)P(B|A_k)}}$$

To compute the probability of any hypothesis after observing evidence B, only need to know:

For all j:

- $P(A_j)$ prior probability of hypotheses A_j
- $P(B|A_j)$ likelihood of evidence under hypothesis A_j

Example

- One fair and one biased coin (0.75 probability heads)
- Select coin at random and flip many times

Problem: compute probability selected coin is biased

Exercise: MATLAB demo + guess posterior

Calculation

Observe HHTHT. What is probability coin is biased?

$$P(\text{fair}) = P(\text{biased}) = \tfrac{1}{2}$$
$$P(\text{HHTHT}|\text{fair}) = \left(\tfrac{1}{2}\right)^5$$
$$P(\text{HHTHT}|\text{biased}) = \left(\tfrac{3}{4}\right)^3\left(\tfrac{1}{4}\right)^2$$

$$P(\text{biased}|\text{HHTHT}) = \frac{P(\text{biased})P(\text{HHTHT}|\text{biased})}{P(\text{biased})P(\text{HHTHT}|\text{biased}) + P(\text{fair})P(\text{HHTHT}|\text{fair})}$$
$$= \frac{\frac{1}{2}\cdot\left(\frac{3}{4}\right)^3\left(\frac{1}{4}\right)^2}{\frac{1}{2}\cdot\left(\frac{3}{4}\right)^3\left(\frac{1}{4}\right)^2 + \frac{1}{2}\cdot\left(\frac{1}{2}\right)^5}$$

Bayesian Classifiers

Observe vector of features $\mathbf{x}$

Predict class $y \in \{0, 1, \ldots, C\}$ with highest probability given features

$$y_{\text{pred}} = \operatorname{argmax}_y p(y|\mathbf{x})$$

Census Example

	x_1 Age	x_2 College?	y Vote?	$P(\omega)$
ω_1	< 30	no	no	0.25
ω_2			yes	0.03
ω_3		yes	no	0.04
ω_4			yes	0.02
ω_5	≥ 30	no	no	0.33
ω_6			yes	0.10
ω_7		yes	no	0.10
ω_8			yes	0.13

$$p(\text{vote} = \text{yes}|\text{age} < 30, \text{college} = \text{no}) = \frac{.03}{.03 + 0.25}$$
$$< 0.5$$
$$\implies \text{predict vote} = \text{no}$$

Aside: Random Variables

Discrete **random variable** (RV): mapping from outcome $\omega \in \Omega$ to finite set of values

$$X_1(\omega) \in \{< 30, \geq 30\}$$
$$X_2(\omega) \in \{\text{no}, \text{yes}\}$$
$$Y(\omega) \in \{\text{no}, \text{yes}\}$$

Aside: Joint Distribution

Joint distribution of a set of random variables: table of probabilities for all possible settings of those RVs

x_1 Age	x_2 College?	y Vote?	$p(x_1, x_2, y)$
< 30	no	no	0.25
		yes	0.03
	yes	no	0.04
		yes	0.02
≥ 30	no	no	0.33
		yes	0.10
	yes	no	0.10
		yes	0.13

$$\Omega = \{< 30, \geq 30\} \times \{\text{no}, \text{yes}\} \times \{\text{no}, \text{yes}\}$$

Notation

Write RV as X instead of $X(\omega)$ when it is understood that X maps from outcomes to values

Shorthand for joint distributions

$$\begin{aligned} p(x_1, x_2, y) &:= P(X_1 = x_1, X_2 = x_2, Y = y) \\ p(y|x) &:= P(Y = y|X = x) \\ p(\mathbf{x}) &:= P(X_1 = x_1, \ldots, X_n = x_n) \end{aligned}$$

And so on... (notation sometimes problematic, but we won't worry about this...)

Bayesian Classifiers

$$\begin{aligned} y_{\text{pred}} &= \text{argmax}_y\, p(y|\mathbf{x}) \\ &= \text{argmax}_y\, \frac{p(y)p(\mathbf{x}|y)}{p(\mathbf{x})} && \text{Bayes rule} \\ &= \text{argmax}_y\, p(y)p(\mathbf{x}|y) && \text{drop denominator} \end{aligned}$$

Need to know $p(y)$, $p(\mathbf{x}|y)$ for each class

Example. Discuss training

Problem

$p(\mathbf{x}|y)$ may be too big to represent or estimate

Example: text classification

- $\mathbf{x} = (x_1, \ldots, x_{5000})$
- x_j: does word j appear in document?
- 2^{5000} table entries to store $p(x_1, \ldots, x_{5000}|y = 1)$
- Similarly impossible to estimate

Naive Bayes

Assume features are independent given class:

$$\begin{aligned} p(x_1, \ldots, x_n|y) &= p(x_1|y)p(x_2|y)\ldots p(x_n|y) \\ &= \prod_{i=1}^{n} p(x_i|y) \end{aligned}$$

Predict:

$$y_{\text{pred}} = \text{argmax}_y\, p(y) \prod_{j=1}^{n} p(x_j|y)$$

Training

Given: training examples $(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})$, need to estimate

- Class priors:

$$p(y = 0), p(y = 1), \ldots, p(y = C)$$

- Class-conditional distribution of feature x_j

$$\begin{aligned} &p(x_j = 0 \,|\, y = c) \\ &p(x_j = 1 \,|\, y = c) \\ &p(x_j = 2 \,|\, y = c) \\ &\ldots \\ &p(x_j = k \,|\, y = c) \end{aligned}$$

($C = \#$ classes; $k = \#$ values of x_j)

Training: Class Prior

Class priors:

$$p(y = c) = \frac{\sum_{i=1}^{m} \mathbf{1}\{y^{(i)} = c\}}{m}$$

(fraction of training examples with class c)

Example

Training: Class-conditional Distribution

Conditional probability that $x_j = v$ given class c:

$$p(x_j = v \mid y = c) = \frac{\sum_{i=1}^m \mathbf{1}\{x_j^{(i)} = v, y^{(i)} = c\}}{\sum_{i=1}^m \mathbf{1}\{y^{(i)} = c\}}$$

(Fraction of examples with $x_j = v$ among those in class c)

Example

Laplace Smoothing

Conditional probability that $x_j = v$ given class c:

$$p(x_j = v \mid y = c) = \frac{1 + \sum_{i=1}^m \mathbf{1}\{x_j^{(i)} = v, y^{(i)} = c\}}{k + \sum_{i=1}^m \mathbf{1}\{y^{(i)} = c\}}$$

(Avoid zero probabilities: pretend there is an extra training example of each type)

Example

Additional Topics

- Discretization of continuous features
- Variations of Naive Bayes for text