CS 335: Probabiilty Review, Bayesian Reasoning, Naive Bayes

Dan Sheldon

Probability Review

Motivation

Age	College?	Vote?	probability
< 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

- Suppose we want to predict whether someone will vote or not given demographic variables. E.g., will a 37-year-old with college degree vote?
- One way to do this is by reasoning about *probabilities* of different combinations of variables
- Informally, probability = frequency in the population

Probability Space

	Age	College?	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

- A **sample space** Ω is a set of possible **outcomes**. We will assume $\Omega = \{\omega_1, \ldots, \omega_n\}$ is discrete and finite.
- Each outcome ω is assigned a **probability** $P(\omega)$. Probabilities are non-negative and sum to one.
  - $P(\omega) \geq 0$
  - $P(\omega_1) + \ldots + P(\omega_n) = 1$

Events

	Age	College?	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

- An **event** $A \subseteq \Omega$ is a subset of the sample space. The probability of A is the sum of probabilities of outcomes in

$$P(A) = \sum_{\omega \in A} P(\omega)$$

- **Example**: "less than 30"
  - $A = \{\omega_1, \omega_2, \omega_3, \omega_4\}$
  - $P(A) = 0.25 + 0.03 + 0.04 + 0.02 = 0.34$
- Less than 30 and college educated?

Events

	Age	College?	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

- **Events are the only things that have probabilities**
- Seemingly informal statements like $P(\leq 30)$, $P(\leq 30$ and voted) are made precise by interpreting the phrases inside $P(\cdot)$ as events
- How would you formalize $P($I will get a haircut tomorrow)?

Joint and Conditional Probability

	Age	College?	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

- The **joint probability** $P(A, B)$ of two events A and B is the probability they both occur:

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

- What is $P(\text{college} = \text{no}, \text{vote} = \text{yes})$?

$$P(\omega_2) + P(\omega_6) = 0.13$$

Joint and Conditional Probability

	Age	College?	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

- The **conditional probability** $P(A \mid B)$ of two events A and B is

$$P(A \mid B) := \frac{P(A, B)}{P(B)}$$

- What is $P(\text{vote} = \text{yes} \mid \text{college} = \text{no})$?

$$\frac{P(\omega_2) + P(\omega_6)}{P(\omega_1) + P(\omega_2) + P(\omega_5) + P(\omega_6)} = \frac{0.03 + 0.10}{0.25 + 0.03 + 0.33 + 0.10} = \frac{0.13}{0.61}$$

Law of Total Probability

	Age	College?	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

- Let $A_1, A_2, \ldots A_k$ be events that *partition* Ω
  - A_i and A_j are disjoint for all $i \neq j$
  - $A_1 \cup A_2 \cup \ldots \cup A_k = \Omega$
- Then, for any other event B

$$P(B) = P(A_1, B) + \ldots + P(A_k, B)$$

- **Example**

$$P(\text{vote} = \text{yes}) = P(\text{vote} = \text{yes}, < 30) + P(\text{vote} = \text{yes}, \geq 30)$$

Bayesian Reasoning

Bayes Rule

Let A and B be two events. Then:

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

(Derivation: apply definition of conditional probability twice)

Interpretation I

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

A = hypothesis

B = evidence

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

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

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

Example:

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

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)

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

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

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

Interpretation II

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

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?

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

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

Bayesian Classification and Naive Bayes

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

	Age	College?	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

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

A Bit More Probability: Random Variables

	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

A **random variable** (RV) is a mapping from outcome $\omega \in \Omega$ to finite set of values

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

We usually just write RV as X instead of $X(\omega)$

Joint Distribution of Random Variables

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

- In ML, our probability space is almost always defined as the **joint distribution of a set of random variables**. We dispense with Ω and ω notation: implicitly defined by RVs
- Outcome = setting of the variables
- Sample space = all possible settings

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

Joint Distribution: Notation

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

Common notation short-hand:

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

Discuss / examples

Bayesian Classifiers

$$y_{\text{pred}} = \operatorname{argmax}_y p(y|\mathbf{x})$$
$$= \operatorname{argmax}_y \frac{p(y)p(\mathbf{x}|y)}{p(\mathbf{x})} \qquad \text{Bayes rule}$$
$$= \operatorname{argmax}_y p(y)p(\mathbf{x}|y) \qquad \text{drop denominator}$$

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

Training Bayesian Classifiers

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

Estimate

- Class priors $p(y = 0), p(y = 1), \ldots, p(y = C)$
- Class-conditional distribution $p(x_1, \ldots, x_n | y = c)$ for *every* joint settting of features $x_1, \ldots, x_n$ and every class c

Problem

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

Example: text classification

- $x_j \in \{0, 1\}$: does word j appear in document?
- 5000 words $\Rightarrow 2^{5000}$ values for $p(x_1, \ldots, x_{5000} | y = 1)$

Naive Bayes

Assume features are *independent* given class:

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

Predict:

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

Need to know $p(y)$, $p(x_j|y)$ for all j. **Much** less information to store/estimate.

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

$$p(x_j = 0 \,|\, y = c)$$
$$p(x_j = 1 \,|\, y = c)$$
$$p(x_j = 2 \,|\, y = c)$$
$$\ldots$$
$$p(x_j = k \,|\, y = c)$$

($C = \#$ 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:

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

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

Example

Laplace Smoothing

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

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

(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