Machine learning and naive Bayes classification units

- Machine learning boot camp (last week Friday)
- Finishing basic ML terms; bag-of-words model (Monday)
- The math of multinomial naive Bayes classification (**Today**)
- Lab: NBC (Friday)
- Practical considerations of NBC (next week Monday)

Today:

- "Bayes"
- "Naive"
- "Multinomial"
- Putting it together

Bayesian vs Frequentist Probability

The frequentist point of view is based on the following postulates:

- F1 *Probability refers to limiting relative frequencies. Probabilities are objective properties of the real world.*
- F2 *Parameters are fixed, unknown constants.*
- F3 *Statistical procedures should be designed to have well-defined long run frequency properties.*

The Bayesian approach is based on the following postulates:

- B1 *Probability describes degree of belief, not limiting frequency. "The probability that Albert Einstein drank a cup of tea on August 1, 1948 is .35" does not refer to any limiting frequency but reflects my strength of belief that the proposition is true.*
- B2 *We can make probabilty statements about parameters, even though they are fixed constants.*
- B3 *We make inferences about a parameter by producing a probability distribution for it.*

Wasserman, All of Statistics, pg 175-176, abridged.

The **conditional probability** of event X in light of event Y is

$$P(X|Y) = \frac{P(XY)}{P(Y)}$$

Bayes's theorem allows us to convert from one conditional probability to another

$$\underbrace{P(X|Y)}_{\text{posterior}} = \frac{\overbrace{P(Y|X)}^{\text{likelihood}} \overbrace{P(X)}^{\text{prior}}}{\underbrace{P(Y)}_{\text{marginal}}}$$

Think of X as the hypothesis and Y as the evidence. What do we think of hypothesis X in light of evidence Y?

Let $\vec{d}$ be a text/document (observation/data point) viewed as a vector of features

$$\vec{d} = \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_{D-1} \end{bmatrix}$$

The **naïve Bayesian assumption** is

$$P(\vec{d}) = P(f_0) \cdot P(f_1) \cdot \ldots \cdot P(f_{D-1})$$

or

$$P(\vec{d}|c) = P(f_0|c) \cdot P(f_1|c) \cdot \ldots \cdot P(f_{D-1}|c)$$

Multinomial distribution

Suppose you have an urn of ball of k colors. Let p_i be the probability of drawing a ball of color i. $\sum_{i=0}^{k-1} p_i = 1$. Assume

$$p_i = \frac{\text{the number of balls of color } i}{\text{the number of balls altogether}}$$

Suppose you draw n times with replacement. Consider the event that you get color 0 x_0 times, color 1 x_1 times, etc, for some x_0, x_1, ... It must be that $\sum_{i=0}^{k-1} x_i = n$. Refer to this event as

$$\vec{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{k-1} \end{bmatrix}$$

The probability of event $\vec{x}$ is

$$f(\vec{x}) = \binom{n}{x_0 x_1 \dots x_{k-1}} p_0^{x_0} p_1^{x_1} \cdots p_{k-1}^{x_{k-1}}$$

Based on Wasserman, pg 39

The multinomial naïve Bayes classifier

To classify text/document/data point d represented as a vector with features $f_0, f_1, \ldots,$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} \; P(c) \cdot \prod P(f_i \mid c)$$

Compare Eq. B.8 in J & M

Rewritten: Let V be the set of word types we care about, let $D = |V|$, let f_i be the count of word/feature i in document d, and let $P(v_j \mid c)$ be the probability of v_j occurring in a doucment of class c. Then

$$\begin{aligned} c_{NB} &= \operatorname{argmax}_{c \in C} \; P(c) \cdot \textstyle\prod_{i=0}^{D-1} P(v_i \mid c)^{f_i} \\ &= \operatorname{argmax}_{c \in C} \; \log P(c) + \textstyle\sum_{i=0}^{D-1} f_i \log P(v_i \mid c) \end{aligned}$$

Coming up:

- Do bag-of-words programming assignment (Wed, Oct 25)
- Read J&M 4.(0-8, 10) (Wed, Oct 25)
- Take NBC quiz (Tues, Nov 4)
- Do NBC programming assignment (Mon, Nov 10)

(There will be some sort of reading for the next unit (stylometry), but it will be from something outside our textbook. I haven't decided on it yet, though.)