LAB 3: NAIVE BAYES CLASSIFIERS, MAXIMUM LIKELIHOOD ESTIMATION

Nikhil Krishnaswamy
Brandeis University
COSI114: Spring 2016

WHAT'S IN A CATEGORY?

Mary Morstan from *Sherlock*, played by Amanda Abbington, © BBC

Bayesian deductive reasoning:

- Category *C* with features *X*
- Object *O* with features *Y*
- How likely that *O* belongs to *C*?

How can you tell if some random woman is a linguist?

Female linguists:
80% wear dangly earrings
60% wear fur coats
75% have short hair
1% of all women

Female cartographers:
10% wear dangly earrings
50% wear fur coats
95% have short hair
5% of all women

WHAT'S IN A CATEGORY?

Female linguists:
P(F1): 80% wear dangly earrings
P(F2:) 60% wear fur coats
P(F3): 75% have short hair
P(C): 1% of all women

Female cartographers:
P(F1): 10% wear dangly earrings
P(F2): 50% wear fur coats
P(F3): 95% have short hair
P(C): 5% of all women

Bayes' Theorem: *posterior = (prior x likelihood) / evidence*

$$P(C|F1,F2,F3) = [P(C) \times P(F1,F2,F3|C)] / P(F1,F2,F3)$$

$$P(F1,F2,F3) = [P(C) \times P(F1|C) \times P(F2|C) \times P(F3|C)] + [P(C') \times P(F1|C') \times P(F2|C') \times P(F3|C')]$$

WHAT'S IN A CATEGORY?

Female linguists:
P(F1): 80% wear dangly earrings
P(F2:) 60% wear fur coats
P(F3): 75% have short hair
P(C): 1% of all women

Female cartographers:
P(F1): 10% wear dangly earrings
P(F2): 50% wear fur coats
P(F3): 95% have short hair
P(C): 5% of all women

P(*Linguist*|Appearance) = [.01 × (.8x.65x.75)] / (.01x.8x.65x.75 + .05x.1x.5x.95) ≈ **62.15%**

P(*Cartographer*|Appearance) = [.05 × (.1x.5x.05)] / (.01x.8x.65x.75 + .05x.1x.5x.95) ≈ 37.85%

BERNOULLI VARIABLES

Bernoulli variable: variable with two possible outcomes
E.g. "yes"/"no", "heads"/"tails", "spam"/"not-spam"

$B(\theta) \sim P(X|\theta)$
How to find θ? M(aximum) L(ikelihood) E(stimate)

Likelihood: probability of known events (Probability ≈ Likelihood in the future)

Bernoulli assumption: all events independent
$P(E1, E2, E3|\theta) = P(E1|\theta)*P(E2|\theta)*P(E3|\theta)$

GET SPAM

Given: Random sample $X_1, X_2, ..., X_n$ where
$X_i = 0$ if e-mail is not spam, $X_i = 1$ if e-mail is spam
Find MLE of θ, the proportion of e-mails that are spam

Data: 3 e-mails, 2 spam, 1 not (3 observations)

MLE = Maximize product ($\prod$) from 1 to n(umber of observations) of P(X = desired observation)

For this data:
$(P(X=1))^2(P(X=0))^1 = \theta^2*(1-\theta)^1$

GET SPAM

For this data:
$(P(X=1))^2(P(X=0))^1 = \theta^2 * (1-\theta)^1$

Maximize: $\theta^2 * (1-\theta)^1$

Use Calculus!
Screw calculus, use Wolfram Alpha!
http://www.wolframalpha.com/input/?i=maximize+p%5E2(1-p)%5E1

GET SICK

Given: Random sample $X_1, X_2, ..., X_n$ where
$X_i = 0$ if patient is not sick, $X_i = 1$ if patient is sick
Find MLE of θ, the proportion of patients that are sick

100 observations: 19 sick, 81 healthy

For this data:
$(P(X=1))^{19}(P(X=0))^{81} = \theta^{19} * (1-\theta)^{81}$

http://www.wolframalpha.com/input/?i=maximize+p%5E19(1-p)%5E81

GET POLITICAL

Given: Random sample $X_1, X_2, ..., X_n$ where
$X_i = 0$ if NSA thinks you're not a terrorist based on data they got off your phone, $X_i = 1$ if they do
Find MLE of θ, the proportion of people NSA flags as terrorists

1 trillion observations: 10,000,000 suspected terrorists, 990,000,000 not

For this data:
$(P(X=1))^{10\text{^}7}(P(X=0))^{9.9\times 10\text{^}7} = \theta^{10\text{^}7} * (1-\theta)^{9.9\times 10\text{^}7}$

I'm not going to run this through Wolfram Alpha, are you crazy?

WHAT GOOD DOES THIS DO?

Allows you to predict future events based on past sample

Principles of the Bernoulli model apply across non-Bernoulli distributions

- E.g. I have the weights of ten people - I can predict the mean weight of all people from this sample
- Obviously larger sample sizes allow better predictions (https://onlinecourses.science.psu.edu/stat414/node/191)

MLE central to tuning feature weights in a Maximum Entropy model or classifier

Statistics don't lie (much) given a well-trained model and a large enough sample size

PYTHON THEN GO HOME