Naive Bayes

Nupur Chatterji, Kenny Marino, Colin White

Outline

- Bayes’ Rule
- Naive Bayes
- MLE/MAP
- Questions about HW1?

Bayes' Rule

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

- P(A) is known as the “prior”
- P(A|B) is known as the “posterior”

Other Forms of Bayes' Rule

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

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

- Law of Total Probability

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

Joint Distributions

We want to know

P(A|T, M, C)

One solution:

Consider all combos of 2^3 features separately

Temp (T)	Mood (M)	Go To Class (C)	Gets A
High	Happy	Yes	.25
High	Happy	No	.05
High	Sad	Yes	.05
High	Sad	No	.15
Low	Happy	Yes	.20
Low	Happy	No	.05
Low	Sad	Yes	.15
Low	Sad	No	.10

Joint Distributions

- Learning P(Y|X) takes way too much data

Solution: Naive Bayes assumption

$$P(X_1, X_2, \ldots, X_n | Y) = \prod_i P(X_i | Y)$$

X_i, X_j are conditionally independent given Y

Conditional Independence

- Height is not independent of weight
- Height is independent of weight, given age

X is conditionally independent of Y given Z if

For all x,y,z, $P(X=x \mid Y=y, Z=z) = P(X=x \mid Z=z)$

Naive Bayes Algorithm

Given training data, features X_1, ...,X_n, label Y

Given a new datapoint $X^{new}=X_1^{new},...,X_n^{new}$

Want to compute most probable label

$Y^{new} = \text{argmax}_y\ P(Y=y|X_1^{new},...,X_n^{new})$

$$Y^{new} = \text{argmax}_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

Naive Bayes Algorithm (Binary features & label)

Given training data, feats $X_1, \ldots, X_n$, label Y

- Estimate $\pi_1 = P(Y=1)$
- For each x_i estimate $\theta_{i,1} = P(X_i = 1 \mid Y=1)$

Given test set,

- Predict 1 if $\pi_1 \Pi_i \theta_i > 0.5$

Maximum Likelihood Estimate

Estimate π_1

- $P(Y=1) = (\# Y=1)/n$

Estimate θ_i

- $P(X_i = 1 \mid Y = 1) = (\# X_i=1 \text{ and } Y = 1) / (\# Y = 1)$

Maximum A Priori

- Recall the prediction rule $\pi_1 \Pi_i \, \theta_i > 0.5$
- If some θ_i = P(X_i = 1 | Y=1) is zero, then we would predict Y=0 for *all* datapoints s.t. X_i = 1 (why?)

Solution: use MAP (hallucinated examples)

- θ_i = [(# X_i=1 and Y = 1)+c] / [(# Y = 1)+2c]

Naive Bayes Assumption

- Very likely, the Naive Bayes assumption is not true

$$P(X_1, X_2, \dots, X_n | Y) = \prod_i P(X_i | Y)$$

- Still, Naive Bayes does well in real life
- Examples where it is extremely not satisfied vs. almost satisfied?

Any other questions?