Perceptron, MLE, MAP

Kenny Marino, Colin White,
Nupur Chatterji

Outline

- Example
- Bayes' Rule
- MLE
- MAP
- Common Mistakes on HW 0
- Questions?

Perceptron Algorithm

- Set t=1, start with the all zero vector w_1
- Given example x, predict positive iff $w_t \cdot x \geq 0$
- On a mistake, update as follows:
  - Mistake on positive, then update $w_{t+1} \leftarrow w_t + x$
  - Mistake on negative, then update $w_{t+1} \leftarrow w_t - x$

Perceptron Algorithm: Example

Point	Label
(1, 2, -1)	+
(2, 0, 0)	-
(1, 3, 4)	+
(0, -2, -1)	-

Bayes' Rule

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

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

Other Forms of Bayes' Rule

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

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

- Law of Total Probability

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

Using Bayes' Rule

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

A = you finished your homework

B = you are tired

P(A) = 0.3

P(B | A) = 0.55

P(B | ~A) = 0.45

What is P(you finished your homework | you are tired) = P(A | B)?

Motivation + Problem + Solution

- Learning $P(Y|X)$ instead of $F: X \rightarrow Y$
  - Joint Distribution Tables
- Requires more data than we have available
- Solution
  - Estimate probabilities from sparse data **SMARTLY**
  - Maximum Likelihood Estimates (MLE)
  - Maximum A Posteriori Estimates (MAP)

Random Variables and Joint Distributions

Temperature	Mood	Go To Classes	Probability
High	Happy	Yes	0.25
High	Happy	No	0.024
High	Sad	Yes	0.042
High	Sad	No	0.012
Low	Happy	Yes	0.33
Low	Happy	No	0.097
Low	Sad	Yes	0.13
Low	Sad	No	0.115

Joint Distribution

- With M variables, we have 2^M rows
- Inference (making a query)
  - Sum over rows (the probabilities) that match the event

Main Ideas

- MLE
  - Choose parameters θ that maximize $P(\text{Data} \mid \theta)$
- MAP
  - Choose parameters θ that maximize $P(\theta \mid \text{Data})$

Maximum Likelihoood Estimate (MLE)

- Assume i.i.d
  - What does this mean?
  - What does this allow us to do?
- We saw an example with Bernoulli Variables in class
- Look at Binomial Variables
  - B(n, p)
  - n = number of trials
  - p = probability of success

Binomial MLE

- $p(X = k \mid p) = C(n,k)\ p^k (1 - p)^{n-k}$
- $\text{argmax}_p\ L(p) = \text{argmax}_p\ [p^k (1 - p)^{n-k}]$
- $\text{argmax}_p\ L(p) = \text{argmax}_p [k\log(p) + (n-k)\log(1-p)]$
- Setting derivative to 0
- $p = k/n$

High Probability Bound

Hoeffding Inequality:

For any $\epsilon > 0$, $\quad P(|\hat{\theta}_{MLE} - \theta| \geq \epsilon) \leq 2\, e^{-2n\epsilon^2}$

High Probability Bound: Want to know the coin parameter θ within $\epsilon > 0$ with probability at least $1 - \delta$. How many flips?

Set $P(|\hat{\theta}_{MLE} - \theta| \geq \epsilon) \leq 2\, e^{-2n\epsilon^2} \leq \delta$ Solve for n: $n \geq \frac{\ln\frac{2}{\delta}}{2\,\epsilon^2}$

Maximum A Priori Estimate (MAP)

- Used when we have some sort of prior knowledge of the data
- Choose parameters θ that are the most probable given observed data and prior beliefs

Binomial MAP

- Let $p(p \mid a, b) = 1/B(a, b)\ p^{a-1}\ (1 - p)^{b-1}$
- $p(X = k \mid p) = C(n,k)\ p^{k}\ (1 - p)^{n-k}$
- $\text{argmax}_p\ p(D \mid p)p(p) = \text{argmax}_p\ \log[p(D \mid p)p(p)]$
- $\text{argmax}_p\ p(D \mid p)p(p) = \text{argmax}_p[(k+a-1)\log(p) + (n-k+b-1)\log(1-p)]$
- Setting derivative to 0
- $p = (k+a-1)/(n+a+b-2)$

Any other questions?