CS 112: Modeling Uncertainty in Information Systems

Prof. Jenn Wortman Vaughan

May 16, 2012

Lecture 13

Reminders & Announcements

- Homework 3 is due this Friday
- We will cover the algorithm that you will implement for Homework 4 in class on Monday

Inference

Probability Theory

- So far, we have assumed the existence of a fully-specified probabilistic model that obeys the axioms of probability.

Probability Theory

- So far, we have assumed the existence of a fully-specified probabilistic model that obeys the axioms of probability.

- The questions that we asked have had a unique right answer with respect to the model.

Probability Theory

- So far, we have assumed the existence of a fully-specified probabilistic model that obeys the axioms of probability.

- The questions that we asked have had a unique right answer with respect to the model.

  - A fair die is rolled three times. What is the probability that all rolls are greater than three?

Probability Theory

- So far, we have assumed the existence of a fully-specified probabilistic model that obeys the axioms of probability.

- The questions that we asked have had a unique right answer with respect to the model.

  - A fair die is rolled three times. What is the probability that all rolls are greater than three?
  - If the time before a hard disk fails is modeled as an exponential random variable with mean λ, how likely is it to fail in the first two years?

Statistical Inference

- In statistical inference, we are given only observations.

Statistical Inference

- In statistical inference, we are given only observations.

- There may not always be a single “right” answer…

Statistical Inference

- In statistical inference, we are given only observations.

- There may not always be a single "right" answer…
  - Based on a collection of old email, how likely is it that this new email is spam?

Statistical Inference

- In statistical inference, we are given only observations.

- There may not always be a single "right" answer…
  - Based on a collection of old email, how likely is it that this new email is spam?

- For the next two classes, we will discuss techniques that can be used to answer questions like this.

Types of Inference

Hypothesis testing: Decide which of two or more hypotheses is most likely to true based on some data.

- Determine whether an email containing a particular set of words is more likely to be spam or not spam
- Given a student's test score, decide if he studied or not

Types of Inference

Hypothesis testing: Decide which of two or more hypotheses is most likely to true based on some data.

- Determine whether an email containing a particular set of words is more likely to be spam or not spam
- Given a student's test score, decide if he studied or not

Parameter estimation: Have a model that is fully specified except some unknown parameters we need to estimate.

- Estimate the bias of a coin from a sequence of flips
- Estimate the fraction of the population who prefers candidate A to candidate B based on polling data

Hypothesis Testing

Hypothesis Testing

Let D be the event that we observed some particular data

- D = event that I observed an email containing the words “ca$h” and “viagra”

Hypothesis Testing

Let D be the event that we observed some particular data

- D = event that I observed an email containing the words "ca$h" and "viagra"

Let $H_1, \ldots, H_k$ be disjoint and exhaustive events representing hypotheses we are choosing among

- H_1 = event that the email is spam
- H_2 = event that the email is not spam

Hypothesis Testing

Let D be the event that we observed some particular data

- D = event that I observed an email containing the words "ca$h" and "viagra"

Let $H_1, \ldots, H_k$ be disjoint and exhaustive events representing hypotheses we are choosing among

- H_1 = event that the email is spam
- H_2 = event that the email is not spam

What is the most likely hypothesis given the data?

Maximum Likelihood

- Suppose that we know (or can compute) the probability $P(D \mid H_i)$ of observing data D for each hypothesis H_i

Maximum Likelihood

- Suppose that we know (or can compute) the probability $P(D \mid H_i)$ of observing data D for each hypothesis H_i

- The maximum likelihood (ML) hypothesis is the hypothesis that makes the data most likely

$$H^{ML} = \text{argmax}_i \, P(D \mid H_i)$$

Maximum Likelihood

When I take the freeway to work, there is a 60% chance that I hit traffic. When I take back roads, there is a 30% chance that I hit traffic. Suppose I tell you that I hit traffic on the way to work today. What is the maximum likelihood hypothesis regarding the route I took?

One Potential Problem

Suppose I tell you that I only take the freeway to work 5% of the time. Does it still seem most likely that I took the freeway today?

One Potential Problem

Suppose I tell you that I only take the freeway to work 5% of the time. Does it still seem most likely that I took the freeway today?

How can we incorporate this information into our reasoning?

Bayesian Reasoning

- If we know $P(H_i)$ and $P(D \mid H_i)$ for each i, we can use Bayes' rule to compute $P(H_i \mid D)$ for each hypothesis

Bayesian Reasoning

- If we know $P(H_i)$ and $P(D \mid H_i)$ for each i, we can use Bayes' rule to compute $P(H_i \mid D)$ for each hypothesis

- $P(H_i)$ is often referred to as the prior probability of H_i while $P(H_i \mid D)$ is referred to as the posterior probability

Bayesian Reasoning

- If we know $P(H_i)$ and $P(D \mid H_i)$ for each i, we can use Bayes' rule to compute $P(H_i \mid D)$ for each hypothesis

- $P(H_i)$ is often referred to as the prior probability of H_i while $P(H_i \mid D)$ is referred to as the posterior probability

- The posterior probability is a refinement of our prior belief about each hypothesis in light of the observed data

Maximum a Posteriori

- The maximum a posteriori (MAP) hypothesis is the hypothesis with the maximum posterior probability

$$H^{MAP} = \text{argmax}_i \, P(H_i \mid D)$$

Maximum a Posteriori

- The maximum a posteriori (MAP) hypothesis is the hypothesis with the maximum posterior probability

$$H^{MAP} = \text{argmax}_i \, P(H_i \mid D) = \text{argmax}_i \, P(D \mid H_i) \, P(H_i)$$

Maximum a Posteriori

- The maximum a posteriori (MAP) hypothesis is the hypothesis with the maximum posterior probability

$$H^{MAP} = \text{argmax}_i \, P(H_i \mid D) = \text{argmax}_i \, P(D \mid H_i) \, P(H_i)$$

When is this the same as maximum likelihood?

Maximum a Posteriori

When I take the freeway to work, there is a 60% chance that I hit traffic. When I take back roads, there is a 30% chance that I hit traffic. I take the freeway 5% of the time. Suppose I tell you that I hit traffic on the way to work today. What is the MAP hypothesis regarding the route I took?

A Very Quick Exercise…

- You decide to monetize your new website by displaying ads. Visitors to your site are 75% UCLA students, 10% programmers, and 15% members of your immediate family. Students click on your ad with probability 0.1. Programmers click on your ad with probability 0.05. Your family members click on your ad with probability 0.4.

- A visitor comes to your site and clicks on an ad. What is the maximum likelihood hypothesis regarding the visitor type? What is the MAP hypothesis?

Parameter Estimation

Parameter Estimation

Suppose that we would like to estimate the unknown bias p of a coin based on observations of the outcomes $X_1, \ldots, X_n$ of n independent tosses of the coin

Parameter Estimation

Suppose that we would like to estimate the unknown bias p of a coin based on observations of the outcomes $X_1, \ldots, X_n$ of n independent tosses of the coin

(This is just like our polling question…)

Parameter Estimation

Suppose that we would like to estimate the unknown bias p of a coin based on observations of the outcomes $X_1, \ldots, X_n$ of n independent tosses of the coin

(This is just like our polling question…)

We can define analogs of both ML and MAP here

Parameter Estimation

- The maximum likelihood (ML) estimate is the parameter value that makes the data most likely

$$\hat{\theta} = \arg\max_{\theta} P(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n; \theta)$$

Parameter Estimation

- The maximum likelihood (ML) estimate is the parameter value that makes the data most likely

$$\hat{\theta} = \arg\max_{\theta} P(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n; \theta)$$

"parameterized by θ"

Parameter Estimation

- The maximum likelihood (ML) estimate is the parameter value that makes the data most likely

$$\hat{\theta} = \arg\max_{\theta} P(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n; \theta)$$

Parameter Estimation

- The maximum likelihood (ML) estimate is the parameter value that makes the data most likely

$$\hat{\theta} = \arg\max_{\theta} P(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n; \theta)$$

- If $X_1, \ldots, X_n$ are independent observations, then

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} P(X_i = k_i; \theta)$$

Parameter Estimation

- The maximum likelihood (ML) estimate is the parameter value that makes the data most likely

$$\hat{\theta} = \arg\max_{\theta} P(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n; \theta)$$

- If $X_1, \ldots, X_n$ are independent observations, then

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} P(X_i = k_i; \theta)$$

"likelihood"

Parameter Estimation

- The maximum likelihood (ML) estimate is the parameter value that makes the data most likely

$$\hat{\theta} = \arg\max_{\theta} P(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n; \theta)$$

- If $X_1, \ldots, X_n$ are independent observations, then

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} P(X_i = k_i; \theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{n} \log\left(P(X_i = k_i; \theta)\right)$$

Parameter Estimation

- The maximum likelihood (ML) estimate is the parameter value that makes the data most likely

$$\hat{\theta} = \arg\max_{\theta} P(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n; \theta)$$

- If $X_1, \ldots, X_n$ are independent observations, then

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} P(X_i = k_i; \theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{n} \log\left(P(X_i = k_i; \theta)\right)$$

"log likelihood"

Parameter Estimation

Suppose that we would like to estimate the unknown bias p of a coin based on observations of the outcomes $X_1, \ldots, X_n$ of n independent tosses of the coin

What is the maximum likelihood estimate?

Maximum Likelihood is Consistent

Consistency: If θ is the true value of the parameter and θ_n is the maximum likelihood estimate after n observations, then for any $\varepsilon > 0$,

$$\lim_{n \to \infty} P\left(\left|\theta_n - \theta\right| \geq \varepsilon\right) = 0$$

Maximum Likelihood is Consistent

Consistency: If θ is the true value of the parameter and θ_n is the maximum likelihood estimate after n observations, then for any $\varepsilon > 0$,

$$\lim_{n \to \infty} P\left(\left|\theta_n - \theta\right| \geq \varepsilon\right) = 0$$

Translation: As the number of observations gets large, the maximum likelihood estimate gets closer and closer to the true parameter value – clearly desirable for an estimate.

MAP Parameter Estimation

- We can define an analog of MAP for parameter estimation too, though we won't go into the details
- Can be useful if the amount of data we have observed is relatively small
- Requires that we have a prior probability distribution over values of the parameter θ