CS 112: Computer System Modeling Fundamentals

Prof. Jenn Wortman Vaughan

May 17, 2011

Lecture 14

Reminders & Announcements

- Any midterm re-grade requests must be submitted in writing by Friday

- New ruby script on courseweb can be used to get extra Twitter data for your classifier, e.g., run

  > ruby getTweets.rb trump en de fr es

  to get recent tweets on Trump in English, German, French, and Spanish

Today…

- More about parameter estimation
  - Using maximum likelihood
  - Using MAP

- Next time: graphical models

Hypothesis Testing

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

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

Hypothesis Testing

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

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

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

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

ML Parameter Estimation

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

$$\Theta^{ML} = \arg\max_{\theta} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \theta)$$

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

$$\Theta^{ML} = \arg\max_{\theta} \prod_{i=1}^{n} P(X_i = x_i; \theta)$$

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

ML Parameter Estimation

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

$$\Theta^{ML} = \arg\max_{\theta} f_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n;\ \theta)$$

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

$$\begin{aligned}\Theta^{ML} &= \arg\max_{\theta} \prod_{i=1}^{n} f_{X_i}(x_i;\ \theta)\\ &= \arg\max_{\theta} \sum_{i=1}^{n} \log\left(f_{X_i}(x_i;\ \theta)\right)\end{aligned}$$

Parameter Estimation

Suppose that the time it takes for a certain model of hard drive to fail is an exponential random variable

We would like to estimate the unknown parameter λ based on n independent observations $X_1, \ldots, X_n$

What is the maximum likelihood estimate?

Log Likelihood for Computation

- The log likelihood has a computational benefit too…

The Good and the Bad of ML

- Maximum likelihood is consistent – as the number of observations gets large, the maximum likelihood estimate gets closer and closer to the true parameter value

The Good and the Bad of ML

- Maximum likelihood is consistent – as the number of observations gets large, the maximum likelihood estimate gets closer and closer to the true parameter value

- But what if we don't have much data?

The Bayesian Point of View

- Instead of treating parameters as fixed but unknown values θ, Bayesians treat them as random variables Θ

The Bayesian Point of View

- Instead of treating parameters as fixed but unknown values θ, Bayesians treat them as random variables Θ

- Can then define the notions of prior and posterior…

The Bayesian Point of View

- Instead of treating parameters as fixed but unknown values θ, Bayesians treat them as random variables Θ

- Can then define the notions of prior and posterior…
  - Prior: $P(\Theta = \theta)$

The Bayesian Point of View

- Instead of treating parameters as fixed but unknown values θ, Bayesians treat them as random variables Θ

- Can then define the notions of prior and posterior…
  - Prior: $P(\Theta = \theta)$ $f_{\Theta}(\theta)$

The Bayesian Point of View

- Instead of treating parameters as fixed but unknown values θ, Bayesians treat them as random variables Θ

- Can then define the notions of prior and posterior…

  - Prior: $P(\Theta = \theta)$ $f_{\Theta}(\theta)$

  - Posterior:

$$P(\Theta = \theta \mid X_1 = x_1, \ldots, X_n = x_n)$$

The Bayesian Point of View

- Instead of treating parameters as fixed but unknown values θ, Bayesians treat them as random variables Θ

- Can then define the notions of prior and posterior…
  - Prior: $P(\Theta = \theta)$ $f_{\Theta}(\theta)$
  - Posterior:

$$P(\Theta = \theta \mid X_1 = x_1, \ldots, X_n = x_n) \qquad f_{\Theta \mid X_1, \ldots, X_n}(\theta \mid x_1, \ldots, x_n)$$

The Bayesian Point of View

- Instead of treating parameters as fixed but unknown values θ, Bayesians treat them as random variables Θ

- Can then define the notions of prior and posterior…
  - Prior: $P(\Theta = \theta)$ $f_{\Theta}(\theta)$
  - Posterior:

$$P(\Theta = \theta \mid X_1 = x_1, \ldots, X_n = x_n) \qquad f_{\Theta \mid X_1, \ldots, X_n}(\theta \mid x_1, \ldots, x_n)$$

- As before, priors may be subjective or estimated from data

MAP Parameter Estimation

- The maximum a posteriori (MAP) estimate is the most likely parameter value given the data

$$\Theta^{MAP} = \arg\max_{\theta} P(\Theta = \theta \mid X_1 = x_1, \ldots, X_n = x_n)$$

MAP Parameter Estimation

- The maximum a posteriori (MAP) estimate is the most likely parameter value given the data

$$\Theta^{MAP} = \arg\max_{\theta} P(\Theta = \theta \mid X_1 = x_1, \ldots, X_n = x_n)$$

$$= \arg\max_{\theta} P(X_1 = x_1, \ldots, X_n = x_n \mid \Theta = \theta) P(\Theta = \theta)$$

MAP Parameter Estimation

- The maximum a posteriori (MAP) estimate is the most likely parameter value given the data

$$\Theta^{MAP} = \arg\max_{\theta} P(\Theta = \theta \mid X_1 = x_1, \ldots, X_n = x_n)$$
$$= \arg\max_{\theta} P(X_1 = x_1, \ldots, X_n = x_n \mid \Theta = \theta) P(\Theta = \theta)$$

- If $X_1, \ldots, X_n$ are independent given Θ, then

$$\Theta^{MAP} = \arg\max_{\theta} P(\Theta = \theta) \prod_{i=1}^{n} P(X_i = x_i \mid \Theta = \theta)$$

MAP Parameter Estimation

- The maximum a posteriori (MAP) estimate is the most likely parameter value given the data

$$\Theta^{MAP} = \arg\max_{\theta} P(\Theta = \theta \mid X_1 = x_1, \ldots, X_n = x_n)$$
$$= \arg\max_{\theta} P(X_1 = x_1, \ldots, X_n = x_n \mid \Theta = \theta) P(\Theta = \theta)$$

- If $X_1, \ldots, X_n$ are independent given Θ, then

$$\Theta^{MAP} = \arg\max_{\theta} P(\Theta = \theta) \prod_{i=1}^{n} P(X_i = x_i \mid \Theta = \theta)$$

- Can use the same log trick here too

MAP Parameter Estimation

- If Θ is continuous, then

$$\Theta^{MAP} = \arg\max_{\theta} f_{\Theta|X_1,\ldots,X_n}(\theta \mid x_1,\ldots,x_n)$$

MAP Parameter Estimation

- If Θ is continuous, then

$$\begin{aligned}\Theta^{MAP} &= \arg\max_{\theta} f_{\Theta|X_1,\ldots,X_n}(\theta \mid x_1,\ldots,x_n) \\ &= \arg\max_{\theta} P(X_1 = x_1,\ldots,X_n = x_n \mid \Theta = \theta) f_{\Theta}(\theta)\end{aligned}$$

MAP Parameter Estimation

- If Θ is continuous, then

$$\Theta^{MAP} = \arg\max_{\theta} f_{\Theta|X_1,\ldots,X_n}(\theta \mid x_1,\ldots,x_n)$$
$$= \arg\max_{\theta} P(X_1 = x_1,\ldots,X_n = x_n \mid \Theta = \theta) f_{\Theta}(\theta)$$

- If X_1, …, X_n are independent given Θ, then

$$\Theta^{MAP} = \arg\max_{\theta} f_{\Theta}(\theta) \prod_{i=1}^{n} P(X_i = x_i \mid \Theta = \theta)$$

MAP Parameter Estimation

- If Θ is continuous, then

$$\Theta^{MAP} = \arg\max_{\theta} f_{\Theta|X_1,\ldots,X_n}(\theta \mid x_1,\ldots,x_n)$$
$$= \arg\max_{\theta} P(X_1 = x_1,\ldots,X_n = x_n \mid \Theta = \theta) f_{\Theta}(\theta)$$

- If $X_1, \ldots, X_n$ are independent given Θ, then

$$\Theta^{MAP} = \arg\max_{\theta} f_{\Theta}(\theta) \prod_{i=1}^{n} P(X_i = x_i \mid \Theta = \theta)$$

- Can define similar estimates for continuous $X_1, \ldots, X_n$

Parameter Estimation

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

Suppose our prior on the parameter is

$$f_\Theta(\theta) = 2 - 4\,|½ - \theta|\,, \qquad \theta \in [0,1]$$

What is the MAP estimate?

Parameter Estimation

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

Suppose our prior on the parameter is

$$f_\Theta(\theta) = 2 - 4\,|½ - \theta|\,, \qquad \theta \in [0,1]$$

What is the MAP estimate?

(we skipped some of the messy details of this derivation in class – see Problem 5, page 446, or try to finish it yourself)

Parameter Estimation

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

Suppose our prior on the parameter is

$$f_\Theta(\theta) = 2 - 4\,|\tfrac{1}{2} - \theta|\,, \qquad \theta \in [0,1]$$

What is the MAP estimate?

How does this compare to the smoothed ML estimate?

The Posterior Distribution

- In addition to using the posterior to calculate MAP estimates, we can also use it to calculate expectations

The Posterior Distribution

- In addition to using the posterior to calculate MAP estimates, we can also use it to calculate expectations

- Example: Consider again our biased coin. Suppose we play the following game:
  - You flip the coin.
  - If the result is heads, you get $1.
  - If the result is tails, you get nothing.

How much profit do you expect to make from this game?