Announcements

HW3

- Out late tonight
- Due Sat 10/9

Quizzes

- Mon 10/4, last 15 min. of class (calculus, optimization, Lagrange)
- Mon 10/11, last 15 min. of class (probability, statistics)

Mathematical Foundations for Machine Learning

Statistics

Instructor: Pat Virtue

Plan

Recitation

Trick coin

- Finding the best parameters
  - Among five different buckets
  - Plotting the log likelihood vs q

Today

Maximum likelihood estimation

Two applications of Bayes rule

Conditional likelihood

Expectation

Variance

Likelihood

Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

$$P(Y = y \mid \underline{\theta})$$

Additional notation

$$p(y \mid \theta)$$

$$p(y ; \theta)$$

$$p_\theta(y)$$

$$p(D \mid \theta)$$

Likelihood and i.i.d

$D = \{y^{(1)}, y^{(2)}, y^{(3)}\}$

Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

i.i.d.: Independent and identically distributed

$$P(Y^{(1)} = y^{(1)}, Y^{(2)} = y^{(2)}, Y^{(3)} = y^{(3)} \mid \theta^{(1)}, \theta^{(2)}, \theta^{(3)})$$

identical

$$P(Y = y^{(1)}, Y = y^{(2)}, Y = y^{(3)} \mid \theta)$$

independent

$$= P(Y = y^{(1)} \mid \theta)\, P(Y = y^{(2)} \mid \theta)\, P(Y = y^{(3)} \mid \theta)$$

Bernoulli Likelihood

$p(\mathcal{D} \mid \theta)$

Bernoulli distribution:

$$Y \sim Bern(\phi) \qquad p(y \mid \phi) = \begin{cases} \phi, & y = 1 \\ 1 - \phi, & y = 0 \end{cases}$$

What is the likelihood for three i.i.d. samples, given parameter ϕ:

$$\mathcal{D} = \{y^{(1)} = 1, y^{(2)} = 1, y^{(3)} = 0\}$$

$$p(\mathcal{D} \mid \phi) = \prod_{i=1}^{N} P(Y = y^{(i)} \mid \phi)$$

$$= \phi \cdot \phi \cdot (1 - \phi)$$

MLE

Estimating Parameters with Likelihood

We model the outcome of a single mysterious weighted-coin flip as a Bernoulli random variable:

$$Y \sim Bern(\phi)$$

$$p(y \mid \phi) = \begin{cases} \phi, & y = 1 \;\; (Heads) \\ 1 - \phi, & y = 0 \;\; (Tails) \end{cases}$$

Given the ordered sequence of coin flip outcomes:

$$[1, 0, 1, 1]$$

What is the estimate of parameter $\hat{\phi}$?

Estimating Parameters with Likelihood

We model the outcome of a single mysterious weighted-coin flip as a Bernoulli random variable:

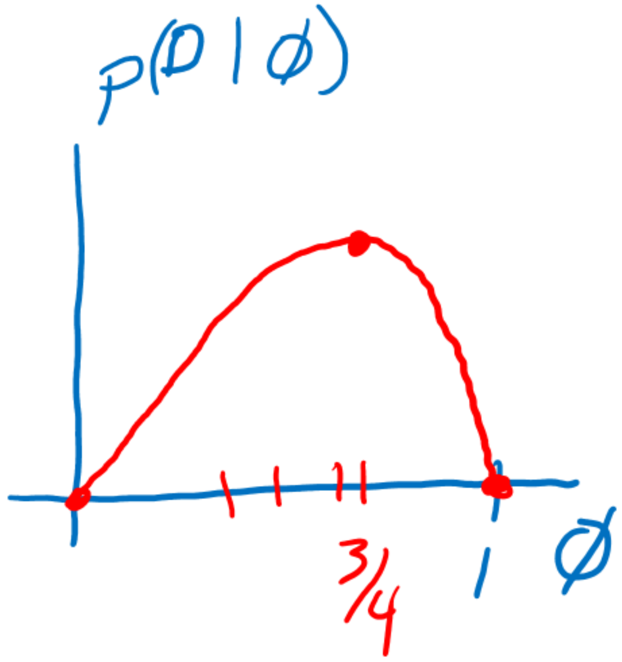
$$Y \sim Bern(\phi)$$

$$p(y \mid \phi) = \begin{cases} \phi, & y = 1 \;\; (Heads) \\ 1 - \phi, & y = 0 \;\; (Tails) \end{cases}$$

Given the ordered sequence of coin flip outcomes:

$$[1, 0, 1, 1]$$

What is the estimate of parameter $\hat{\phi}$?



$$p(\, D \mid \phi \,) = \phi \cdot \phi \cdot (1 - \phi) \cdot \phi$$
$$= \phi^3 (1 - \phi)^1$$

Likelihood and Maximum Likelihood Estimation

Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

Likelihood function: The value of likelihood as we change theta

(same as likelihood, but conceptually we are considering many different values of the parameters)

Maximum Likelihood Estimation (MLE): Find the parameter value that maximizes the likelihood.

MLE as Data Increases

Given the ordered sequence of coin flip outcomes:

$$[1, 0, 1, 1]$$

$$\mathrm{p}(\mathcal{D} \mid \phi) = \prod_{i}^{N} p\left(y^{(i)} \mid \phi\right) = \phi^{N_{y=1}}(1-\phi)^{N_{y=0}}$$

What happens as we flip more coins?

MLE for Gaussian

Gaussian distribution:

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

$$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

What is the log likelihood for three i.i.d. samples, given parameters μ, σ^2?

$$\mathcal{D} = \{y^{(1)} = 65, y^{(2)} = 95, y^{(3)} = 85\}$$

$$L(\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(y^{(i)}-\mu\right)^2}{2\sigma^2}}$$

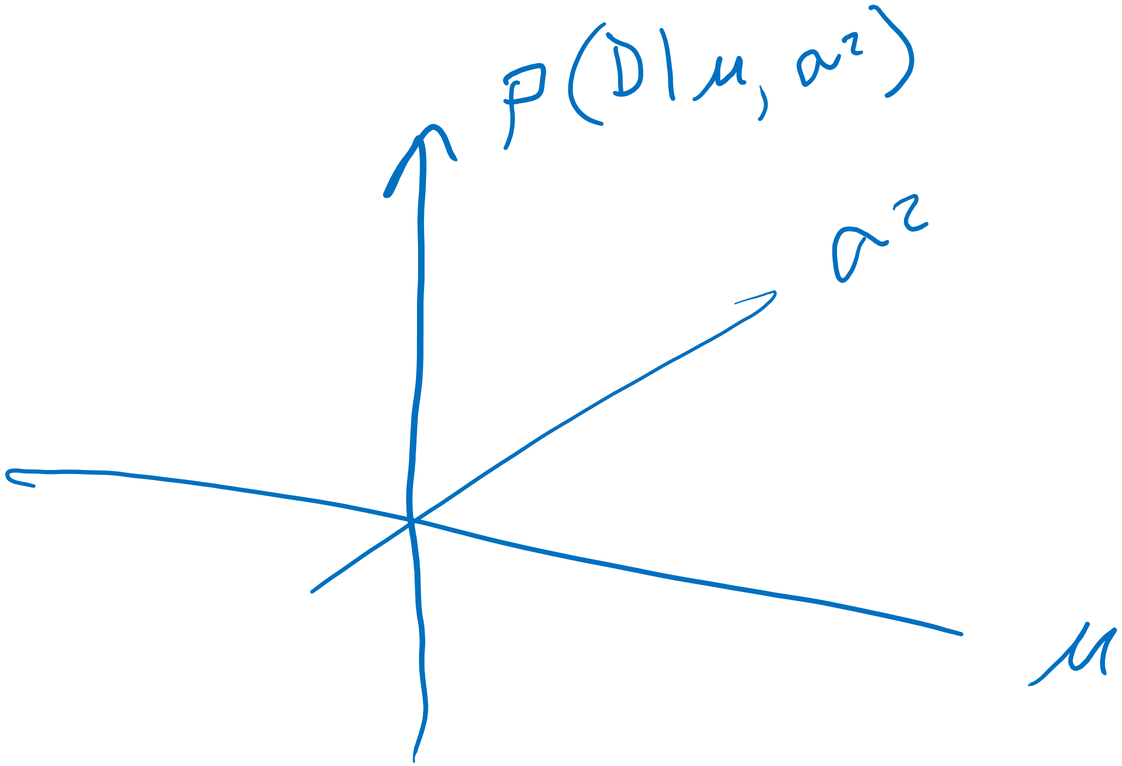
$$\hat{\theta}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i}^{N} p\left(y^{(i)} \mid \boldsymbol{\theta}\right)$$

MLE for Gaussian

Assume that exam scores are drawn independently from the same Gaussian (Normal) distribution.

Given three exam scores 75, 80, 90, which pair of parameters is the best fit (the highest likelihood)?

$$p(\mathcal{D}|\mu,\sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(y^{(i)}-\mu\right)^2}{2\sigma^2}}$$



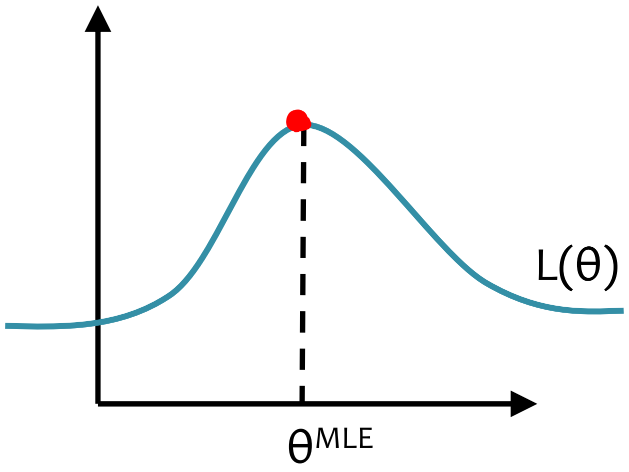
MLE

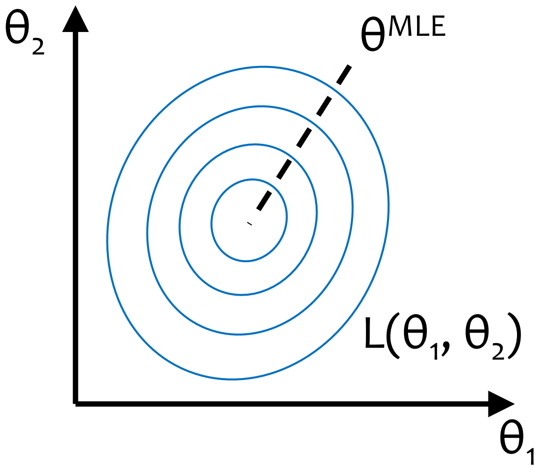
Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

Principle of Maximum Likelihood Estimation:
Choose the parameters that maximize the likelihood of the data.

$$\boldsymbol{\theta}^{\text{MLE}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE)





MLE Recipe

Likelihood and Log Likelihood

$\log xz = \log x + \log z$

Likelihood: The probability (or density) of random variable Y taking on value y given the distribution parameters, $\boldsymbol{\theta}$.

$$p(D \mid \theta) = \prod p(y^{(i)} \mid \theta)$$

Likelihood function: The value of likelihood as we change theta

(same as likelihood, but conceptually we are considering many different values of the parameters)

likelihood $\mathcal{L}(\theta; D) = p(D \mid \theta) = \prod p(y^{(i)} \mid \theta)$

log likelihood $\ell(\theta; D) = \log p(D \mid \theta) = \sum \log p(y^{(i)} \mid \theta)$

Maximum Likelihood Estimation

MLE of parameter θ for i.i.d. dataset $\mathcal{D} = \{y^{(i)}\}_{i=1}^{N}$

$$\hat{\theta}_{MLE} = \operatorname*{argmax}_{\theta} p(\mathcal{D} \mid \theta)$$

Recipe for Estimation

MLE

1. Formulate the likelihood, $p(\mathcal{D} \mid \theta)$
2. Set objective $J(\theta)$ equal to negative log likelihood

$$\mathrm{J}(\theta) = -\log p(\mathcal{D} \mid \theta)$$

3. Compute derivative of objective, $\partial J / \partial \theta$
4. Find $\hat{\theta}$, either
   a. Set derivate equal to zero and solve for θ
   b. Use (stochastic) gradient descent to step towards better θ

MLE for Gaussian

Gaussian distribution:

$$Y \sim \mathcal{N}(\mu, \sigma^2)$$

$$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

What is the log likelihood for three i.i.d. samples, given parameters μ, σ^2?

$$\mathcal{D} = \{y^{(1)} = 65, y^{(2)} = 95, y^{(3)} = 85\}$$

$$L(\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(y^{(i)}-\mu\right)^2}{2\sigma^2}}$$

$$\ell(\mu, \sigma^2) = \sum_{i=1}^{N} -\log\sqrt{2\pi\sigma^2} - \frac{\left(y^{(i)} - \mu\right)^2}{2\sigma^2}$$

$$\hat{\theta}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i}^{N} p\left(y^{(i)} \mid \boldsymbol{\theta}\right)$$

$$\hat{\theta}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i}^{N} \log p\left(y^{(i)} \mid \boldsymbol{\theta}\right)$$

Exercise: MLE for Exponential

Exponential distribution pdf:

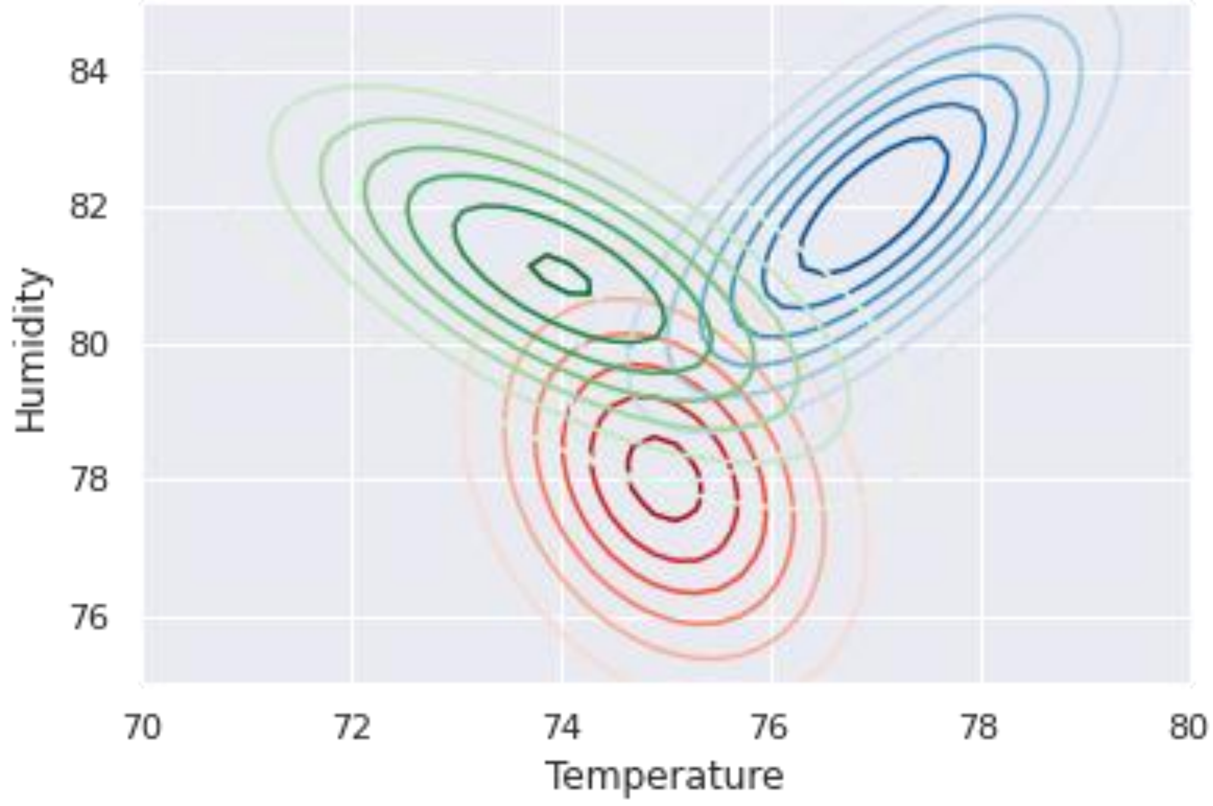
$$f(x) = \lambda e^{-\lambda x}$$

Two Applications of Bayes Rule

Prev Recitation

City classification

There are three nearby cities: City A, City B, and City C. Let $Y \sim Categorical(a, b, c)$ be a categorical distribution where $Y = 1$ means a randomly sampled sensor is in City A, $Y = 2$ means it is in City B, and $Y = 3$ means it is in City C.



Poll 2

Which of these terms is the likelihood?

$$\overset{A}{p(y \mid x)} = \frac{\overset{B}{p(x \mid y)}\ \overset{C}{p(y)}}{\underset{D}{p(x)}}$$

Bayes Rule

Terminology

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}$$

Bayes Rule

Inserting parameters

$$p(\, y \mid x \,) = \frac{p(\, x \mid y \,)\, p(y)}{p(x)}$$

Bayes Rule

Another way to use Bayes rule

$$p(\theta \mid x) = \frac{p(x \mid \theta)\, p(\theta)}{p(x)}$$

Poll 3

Where do we plug in the pdf, $f(x) = \lambda e^{-\lambda x}$

A

B

C

$$\underbrace{p(\theta \mid x)}_{A} = \frac{\overbrace{p(x \mid \theta)}^{B}\ \overbrace{p(\theta)}^{C}}{\underbrace{p(x)}_{D}}$$

D

MLE vs. MAP

Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

$$\boldsymbol{\theta}^{\text{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

Maximum Likelihood Estimate (MLE)

$$\boldsymbol{\theta}^{\text{MAP}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta}) \underbrace{p(\boldsymbol{\theta})}_{\text{Prior}}$$

Maximum *a posteriori* (MAP) estimate

Slide credit: CMU MLD Matt Gormley

Expectation and Variance

Expectation and Variance

The **expected value** of X is $E[X]$. Also called the mean.

- Discrete random variables:

  Suppose X can take any value in the set $\mathcal{X}$.

  $$E[X] = \sum_{x \in \mathcal{X}} x p(x)$$

- Continuous random variables:

  $$E[X] = \int_{-\infty}^{+\infty} x f(x) dx$$

Slide credit: CMU MLD Matt Gormley

Expectation and Variance

The **variance** of X is $Var(X)$.

$$Var(X) = E[(X - E[X])^2]$$

$$\mu = E[X]$$

- Discrete random variables:

$$Var(X) = \sum_{x \in \mathcal{X}} (x - \mu)^2 p(x)$$

- Continuous random variables:

$$Var(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$

Slide credit: CMU MLD Matt Gormley