

Learning with Maximum Likelihood

Note to other teachers and users of these slides. Andrew would be delighted if you found this source material useful in giving your own lectures. Feel free to use these slides verbatim, or to modify them to fit your own needs. PowerPoint originals are available. If you make use of a significant portion of these slides in your own lecture, please include this message, or the following link to the source repository of Andrew's tutorials: <http://www.cs.cmu.edu/~awm/tutorials>. Comments and corrections gratefully received.

Andrew W. Moore
Associate Professor
School of Computer Science
Carnegie Mellon University

www.cs.cmu.edu/~awm
awm@cs.cmu.edu
412-268-7599

Copyright © 2001, Andrew W. Moore

Sep 6th, 2001

Maximum Likelihood learning of Gaussians for Data Mining

- Why we should care
- Learning Univariate Gaussians
- Learning Multivariate Gaussians
- What's a biased estimator?
- Bayesian Learning of Gaussians

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 2

Why we should care

- Maximum Likelihood Estimation is a very very very very fundamental part of data analysis.
- “MLE for Gaussians” is training wheels for our future techniques
- Learning Gaussians is more useful than you might guess...

Learning Gaussians from Data

- Suppose you have $x_1, x_2, \dots, x_R \sim (\text{i.i.d}) N(\mu, \sigma^2)$
- But you don't know μ
(you do know σ^2)

MLE: For which μ is x_1, x_2, \dots, x_R most likely?

MAP: Which μ maximizes $p(\mu | x_1, x_2, \dots, x_R, \sigma^2)$?

Learning Gaussians from Data

- Suppose you have $x_1, x_2, \dots, x_R \sim (\text{i.i.d}) N(\mu, \sigma^2)$
- But you don't know μ

(you do know σ^2)



MLE: For which μ is x_1, x_2, \dots, x_R most likely?



MAP: Which μ maximizes $p(\mu | x_1, x_2, \dots, x_R, \sigma^2)$?

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 5

Learning Gaussians from Data

- Suppose you have $x_1, x_2, \dots, x_R \sim (\text{i.i.d}) N(\mu, \sigma^2)$
- But you don't know μ

(you do know σ^2)



MLE: For which μ is x_1, x_2, \dots, x_R most likely?



MAP: Which μ maximizes $p(\mu | x_1, x_2, \dots, x_R, \sigma^2)$?

Despite this, we'll spend 95% of our time on MLE. Why? Wait and see...

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 6

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \dots, x_R \sim (\text{i.i.d}) N(\mu, \sigma^2)$
- But you don't know μ (you do know σ^2)
- MLE: For which μ is x_1, x_2, \dots, x_R most likely?

$$\mathbf{m}^{mle} = \arg \max_m p(x_1, x_2, \dots, x_R | \mathbf{m}, \mathbf{S}^2)$$

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 7

Algebra Euphoria

$$\mathbf{m}^{mle} = \arg \max_m p(x_1, x_2, \dots, x_R | \mathbf{m}, \mathbf{S}^2)$$

=

(by i.i.d)

=

(monotonicity of log)

=

(plug in formula for Gaussian)

=

(after simplification)

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 8

Intermission: A General Scalar MLE strategy

Task: Find MLE θ assuming known form for $p(\text{Data} | \theta, \text{stuff})$

1. Write $LL = \log P(\text{Data} | \theta, \text{stuff})$
2. Work out $\partial LL / \partial \theta$ using high-school calculus
3. Set $\partial LL / \partial \theta = 0$ for a maximum, creating an equation in terms of θ
4. Solve it*
5. Check that you've found a maximum rather than a minimum or saddle-point, and be careful if θ is constrained

*This is a perfect example of something that works perfectly in all textbook examples and usually involves surprising pain if you need it for something new.

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 9

The MLE μ

$$\mathbf{m}^{mle} = \arg \max_{\mathbf{m}} p(x_1, x_2, \dots, x_R | \mathbf{m}, \mathbf{s}^2)$$

$$= \arg \min_{\mathbf{m}} \sum_{i=1}^R (x_i - \mathbf{m})^2$$

$$= \mathbf{m} \text{ s.t. } 0 = \frac{\partial LL}{\partial \mathbf{m}} =$$

= (what?)

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 10

Lawks-a-lawdy!

$$\mathbf{m}^{mle} = \frac{1}{R} \sum_{i=1}^R x_i$$

- The best estimate of the mean of a distribution is the mean of the sample!

At first sight:

This kind of pedantic, algebra-filled and ultimately unsurprising fact is exactly the reason people throw down their "Statistics" book and pick up their "Agent Based Evolutionary Data Mining Using The Neuro-Fuzz Transform" book.

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 11

A General MLE strategy

Suppose $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n)^T$ is a vector of parameters.

Task: Find MLE \mathbf{q} assuming known form for $p(\text{Data} | \mathbf{q}, \text{stuff})$

1. Write $LL = \log P(\text{Data} | \mathbf{q}, \text{stuff})$
2. Work out $\partial LL / \partial \mathbf{q}$ using high-school calculus

$$\frac{\partial LL}{\partial \mathbf{q}} = \begin{pmatrix} \frac{\partial LL}{\partial \mathbf{q}_1} \\ \frac{\partial LL}{\partial \mathbf{q}_2} \\ \vdots \\ \frac{\partial LL}{\partial \mathbf{q}_n} \end{pmatrix}$$

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 12

A General MLE strategy

Suppose $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n)^\top$ is a vector of parameters.

Task: Find MLE \mathbf{q} assuming known form for $p(\text{Data} | \mathbf{q}, \text{stuff})$

1. Write $LL = \log P(\text{Data} | \mathbf{q}, \text{stuff})$
2. Work out $\partial LL / \partial \mathbf{q}$ using high-school calculus
3. Solve the set of simultaneous equations

$$\begin{aligned}\frac{\partial LL}{\partial \theta_1} &= 0 \\ \frac{\partial LL}{\partial \theta_2} &= 0 \\ &\vdots \\ \frac{\partial LL}{\partial \theta_n} &= 0\end{aligned}$$

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 13

A General MLE strategy

Suppose $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n)^\top$ is a vector of parameters.

Task: Find MLE \mathbf{q} assuming known form for $p(\text{Data} | \mathbf{q}, \text{stuff})$

1. Write $LL = \log P(\text{Data} | \mathbf{q}, \text{stuff})$
2. Work out $\partial LL / \partial \mathbf{q}$ using high-school calculus
3. Solve the set of simultaneous equations

$$\begin{aligned}\frac{\partial LL}{\partial \theta_1} &= 0 \\ \frac{\partial LL}{\partial \theta_2} &= 0 \\ &\vdots \\ \frac{\partial LL}{\partial \theta_n} &= 0\end{aligned}$$

4. Check that you're at a maximum

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 14

A General MLE strategy

Suppose $\mathbf{q} = (\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n)^T$ is a vector of parameters.

Task: Find MLE \mathbf{q} assuming known form for $p(\text{Data} | \mathbf{q}, \text{stuff})$

1. Write $LL = \log P(\text{Data} | \mathbf{q}, \text{stuff})$
2. Work out $\partial LL / \partial \mathbf{q}$ using high-school calculus
3. Solve the set of simultaneous equations

If you can't solve them,
what should you do?

$$\begin{aligned} \frac{\partial LL}{\partial \theta_1} &= 0 \\ \frac{\partial LL}{\partial \theta_2} &= 0 \\ &\vdots \\ \frac{\partial LL}{\partial \theta_n} &= 0 \end{aligned}$$

4. Check that you're at a maximum

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 15

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \dots, x_R \sim (\text{i.i.d}) N(\mu, \sigma^2)$
- But you don't know μ or σ^2
- MLE: For which $\mathbf{q} = (\mu, \sigma^2)$ is x_1, x_2, \dots, x_R most likely?

$$\log p(x_1, x_2, \dots, x_R | \mathbf{m}, \mathbf{s}^2) = -R(\log p + \frac{1}{2} \log \mathbf{s}^2) - \frac{1}{2\mathbf{s}^2} \sum_{i=1}^R (x_i - \mathbf{m})^2$$

$$\frac{\partial LL}{\partial \mathbf{m}} = \frac{1}{\mathbf{s}^2} \sum_{i=1}^R (x_i - \mathbf{m})$$

$$\frac{\partial LL}{\partial \mathbf{s}^2} = -\frac{R}{2\mathbf{s}^2} + \frac{1}{2\mathbf{s}^4} \sum_{i=1}^R (x_i - \mathbf{m})^2$$

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 16

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \dots, x_R \sim (\text{i.i.d}) N(\mu, \sigma^2)$
- But you don't know μ or σ^2
- MLE: For which $\mathbf{q} = (\mu, \sigma^2)$ is x_1, x_2, \dots, x_R most likely?

$$\log p(x_1, x_2, \dots, x_R | \mathbf{m}, \mathbf{s}^2) = -R(\log p + \frac{1}{2} \log \mathbf{s}^2) - \frac{1}{2\mathbf{s}^2} \sum_{i=1}^R (x_i - \mathbf{m})^2$$

$$0 = \frac{1}{\mathbf{s}^2} \sum_{i=1}^R (x_i - \mathbf{m})$$

$$0 = -\frac{R}{2\mathbf{s}^2} + \frac{1}{2\mathbf{s}^4} \sum_{i=1}^R (x_i - \mathbf{m})^2$$

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 17

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \dots, x_R \sim (\text{i.i.d}) N(\mu, \sigma^2)$
- But you don't know μ or σ^2
- MLE: For which $\mathbf{q} = (\mu, \sigma^2)$ is x_1, x_2, \dots, x_R most likely?

$$\log p(x_1, x_2, \dots, x_R | \mathbf{m}, \mathbf{s}^2) = -R(\log p + \frac{1}{2} \log \mathbf{s}^2) - \frac{1}{2\mathbf{s}^2} \sum_{i=1}^R (x_i - \mathbf{m})^2$$

$$0 = \frac{1}{\mathbf{s}^2} \sum_{i=1}^R (x_i - \mathbf{m}) \Rightarrow \mathbf{m} = \frac{1}{R} \sum_{i=1}^R x_i$$

$$0 = -\frac{R}{2\mathbf{s}^2} + \frac{1}{2\mathbf{s}^4} \sum_{i=1}^R (x_i - \mathbf{m})^2 \Rightarrow \text{what?}$$

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 18

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \dots, x_R \sim (\text{i.i.d}) N(\mu, \sigma^2)$
- But you don't know μ or σ^2
- MLE: For which $\mathbf{q} = (\mu, \sigma^2)$ is x_1, x_2, \dots, x_R most likely?

$$\mathbf{m}^{mle} = \frac{1}{R} \sum_{i=1}^R x_i$$

$$\mathbf{s}_{mle}^2 = \frac{1}{R} \sum_{i=1}^R (x_i - \mathbf{m}^{mle})^2$$

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 19

Unbiased Estimators

- An estimator of a parameter is **unbiased** if the expected value of the estimate is the **same** as the true value of the parameters.
- *If* $x_1, x_2, \dots, x_R \sim (\text{i.i.d}) N(\mu, \sigma^2)$ then

$$E[\mathbf{m}^{mle}] = E\left[\frac{1}{R} \sum_{i=1}^R x_i\right] = \mu$$

μ^{mle} is unbiased

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 20

Biased Estimators

- An estimator of a parameter is **biased** if the expected value of the estimate is **different from** the true value of the parameters.
- If $x_1, x_2, \dots, x_R \sim (\text{i.i.d}) N(\mu, \sigma^2)$ then

$$E[\mathbf{s}_{mle}^2] = E\left[\frac{1}{R} \sum_{i=1}^R (x_i - \mathbf{m}^{mle})^2\right] = E\left[\frac{1}{R} \left(\sum_{i=1}^R x_i - \frac{1}{R} \sum_{j=1}^R x_j\right)^2\right] \neq \mathbf{s}^2$$

σ_{mle}^2 is biased

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 21

MLE Variance Bias

- If $x_1, x_2, \dots, x_R \sim (\text{i.i.d}) N(\mu, \sigma^2)$ then

$$E[\mathbf{s}_{mle}^2] = E\left[\frac{1}{R} \left(\sum_{i=1}^R x_i - \frac{1}{R} \sum_{j=1}^R x_j\right)^2\right] = \left(1 - \frac{1}{R}\right) \mathbf{s}^2 \neq \mathbf{s}^2$$

Intuition check: consider the case of $R=1$

Why should our guts expect that σ_{mle}^2 would be an underestimate of true σ^2 ?

How could you prove that?

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 22

Unbiased estimate of Variance

- If $X_1, X_2, \dots, X_R \sim (\text{i.i.d}) N(\mu, \sigma^2)$ then

$$E[\mathbf{s}_{mle}^2] = E\left[\frac{1}{R}\left(\sum_{i=1}^R x_i - \frac{1}{R}\sum_{j=1}^R x_j\right)^2\right] = \left(1 - \frac{1}{R}\right)\mathbf{s}^2 \neq \mathbf{s}^2$$

So define $\mathbf{s}_{\text{unbiased}}^2 = \frac{\mathbf{s}_{mle}^2}{\left(1 - \frac{1}{R}\right)}$ So $E[\mathbf{s}_{\text{unbiased}}^2] = \mathbf{s}^2$

Unbiased estimate of Variance

- If $X_1, X_2, \dots, X_R \sim (\text{i.i.d}) N(\mu, \sigma^2)$ then

$$E[\mathbf{s}_{mle}^2] = E\left[\frac{1}{R}\left(\sum_{i=1}^R x_i - \frac{1}{R}\sum_{j=1}^R x_j\right)^2\right] = \left(1 - \frac{1}{R}\right)\mathbf{s}^2 \neq \mathbf{s}^2$$

So define $\mathbf{s}_{\text{unbiased}}^2 = \frac{\mathbf{s}_{mle}^2}{\left(1 - \frac{1}{R}\right)}$ So $E[\mathbf{s}_{\text{unbiased}}^2] = \mathbf{s}^2$

$$\mathbf{s}_{\text{unbiased}}^2 = \frac{1}{R-1} \sum_{i=1}^R (x_i - \mathbf{m}^{mle})^2$$

Unbiasedness discussion

- Which is best?

$$s_{mle}^2 = \frac{1}{R} \sum_{i=1}^R (x_i - \mathbf{m}^{mle})^2$$

$$s_{unbiased}^2 = \frac{1}{R-1} \sum_{i=1}^R (x_i - \mathbf{m}^{mle})^2$$

Answer:

- It depends on the task
- And doesn't make much difference once $R \rightarrow \text{large}$

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 25

Don't get too excited about being unbiased

- Assume $x_1, x_2, \dots, x_R \sim (\text{i.i.d}) N(\mu, \sigma^2)$
- Suppose we had these estimators for the mean

$$\mathbf{m}^{suboptimal} = \frac{1}{R + 7\sqrt{R}} \sum_{i=1}^R x_i$$

$$\mathbf{m}^{crap} = x_1$$

Are either of these unbiased?

Will either of them asymptote to the correct value as R gets large?

Which is more useful?

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 26

MLE for m-dimensional Gaussian

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R \sim (\text{i.i.d}) N(\mathbf{m}, \mathbf{S})$
- But you don't know \mathbf{m} or \mathbf{S}
- MLE: For which $\mathbf{q} = (\boldsymbol{\mu}, \mathbf{S})$ is $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R$ most likely?

$$\boldsymbol{\mu}^{mle} = \frac{1}{R} \sum_{k=1}^R \mathbf{x}_k$$

$$\mathbf{S}^{mle} = \frac{1}{R} \sum_{k=1}^R (\mathbf{x}_k - \boldsymbol{\mu}^{mle})(\mathbf{x}_k - \boldsymbol{\mu}^{mle})^T$$

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 27

MLE for m-dimensional Gaussian

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R \sim (\text{i.i.d}) N(\mathbf{m}, \mathbf{S})$
- But you don't know \mathbf{m} or \mathbf{S}
- MLE: For which $\mathbf{q} = (\boldsymbol{\mu}, \mathbf{S})$ is $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R$ most likely?

$$\boldsymbol{\mu}^{mle} = \frac{1}{R} \sum_{k=1}^R \mathbf{x}_k$$

$$\mu_i^{mle} = \frac{1}{R} \sum_{k=1}^R x_{ki}$$

Where $1 \leq i \leq m$

And x_{ki} is value of the i^{th} component of \mathbf{x}_k
(the i^{th} attribute of the k^{th} record)

$$\mathbf{S}^{mle} = \frac{1}{R} \sum_{k=1}^R (\mathbf{x}_k - \mathbf{m}^{mle})(\mathbf{x}_k - \mathbf{m}^{mle})^T$$

And μ_i^{mle} is the i^{th} component of \mathbf{m}^{mle}

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 28

MLE for m-dimensional Gaussian

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R \sim (\text{i.i.d}) N(\mathbf{m}, S)$
- But you don't know \mathbf{m} or S
- MLE: For which $\mathbf{q} = (\boldsymbol{\mu}, S)$ is $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R$ most likely?

$$\boldsymbol{\mu}^{mle} = \frac{1}{R} \sum_{k=1}^R \mathbf{x}_k$$

$$\mathbf{S}^{mle} = \frac{1}{R} \sum_{k=1}^R (\mathbf{x}_k - \mathbf{m}^{mle})(\mathbf{x}_k - \mathbf{m}^{mle})^T$$

$$S_{ij}^{mle} = \frac{1}{R} \sum_{k=1}^R (x_{ki} - m_i^{mle})(x_{kj} - m_j^{mle})$$

Where $1 \leq i \leq m, 1 \leq j \leq m$

And x_{ki} is value of the i^{th} component of \mathbf{x}_k (the i^{th} attribute of the k^{th} record)

And σ_{ij}^{mle} is the $(i,j)^{\text{th}}$ component of S^{mle}

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 29

MLE for m-dimensional Gaussian

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \dots$
- But you don't know \mathbf{m} or S
- MLE: For which $\mathbf{q} = (\boldsymbol{\mu}, S)$ is

$$\boldsymbol{\mu}^{mle} = \frac{1}{R} \sum_{k=1}^R \mathbf{x}_k$$

$$\mathbf{S}^{mle} = \frac{1}{R} \sum_{k=1}^R (\mathbf{x}_k - \mathbf{m}^{mle})(\mathbf{x}_k - \mathbf{m}^{mle})^T$$

$$\mathbf{S}^{\text{unbiased}} = \frac{\mathbf{S}^{mle}}{1 - \frac{1}{R}} = \frac{1}{R-1} \sum_{k=1}^R (\mathbf{x}_k - \mathbf{m}^{mle})(\mathbf{x}_k - \mathbf{m}^{mle})^T$$

Q: How would you prove this?

A: Just plug through the MLE recipe.

Note how S^{mle} is forced to be symmetric non-negative definite

Note the unbiased case

How many datapoints would you need before the Gaussian has a chance of being non-degenerate?

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 30

Confidence intervals

We need to talk

We need to discuss how accurate we expect \mathbf{m}^{mle} and \mathbf{S}^{mle} to be as a function of R

And we need to consider how to estimate these accuracies from data...

- Analytically *
- Non-parametrically (using randomization and bootstrapping) *

But we won't. Not yet.

*Will be discussed in future Andrew lectures...just before we need this technology.

Structural error

Actually, we need to talk about something else too..

What if we do all this analysis when the true distribution is in fact not Gaussian?

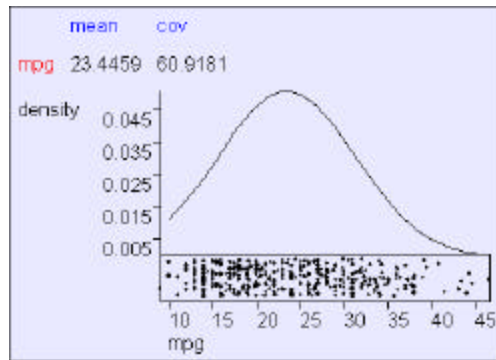
How can we tell? *

How can we survive? *

*Will be discussed in future Andrew lectures...just before we need this technology.

Gaussian MLE in action

Using $N=392$ cars from the "MPG" UCI dataset supplied by Ross Quinlan



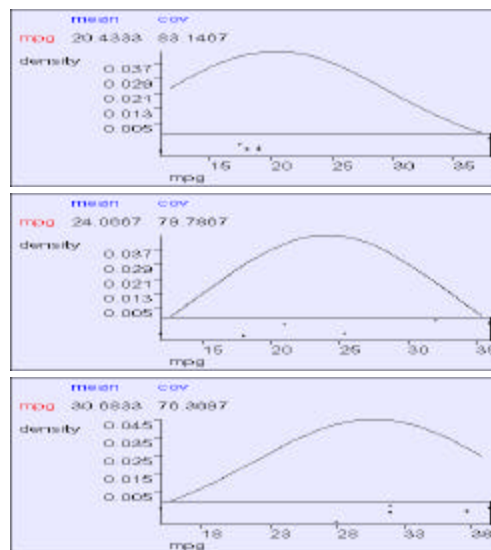
Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 33

Data-starved Gaussian MLE

Using three subsets of MPG.

Each subset has 6 randomly-chosen cars.

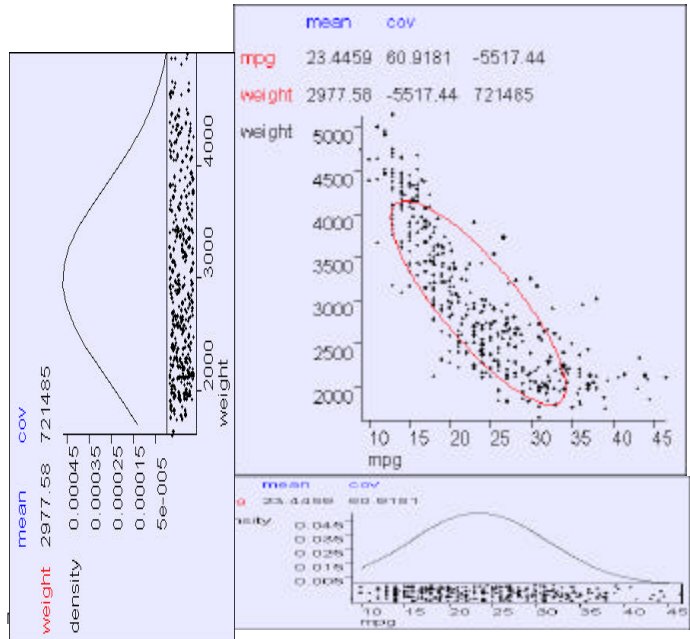


Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 34

Bivariate MLE in action

Copyright © 2001, Andrew W. Moore



Multivariate MLE

	mean	cov
mpg	23.4459	60.9181 -10.3529 -857.585 -233.858 -5517.44 9.11551 18.6915
cylinders	5.47194	-10.3529 2.9097 189.722 55.3482 1300.42 -2.37505 -2.17193
displacement	194.412	-857.585 169.722 10950.4 3614.03 82929.1 -156.994 -142.572
horsepower	104.469	-233.858 55.3482 3614.03 1481.57 28265.6 -73.187 -59.0364
weight	2977.58	-5517.44 1300.42 82929.1 28265.6 721485 -976.815 -987.228
acceleration	15.5413	9.11551 -2.37505 -156.994 -73.187 -976.815 7.61133 2.95046
modelyear	75.9796	18.6915 -2.17193 -142.572 -59.0364 -987.228 2.95046 13.5699

Covariance matrices are not exciting to look at

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 36

Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R \sim (\text{i.i.d}) N(\mathbf{m}, S)$
- But you don't know \mathbf{m} or S
- MAP: Which (\mathbf{m}, S) maximizes $p(\mathbf{m}, S | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R)$?



Step 1: Put a prior on (\mathbf{m}, S)

Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R \sim (\text{i.i.d}) N(\mathbf{m}, S)$
- But you don't know \mathbf{m} or S
- MAP: Which (\mathbf{m}, S) maximizes $p(\mathbf{m}, S | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R)$?



Step 1: Put a prior on (\mathbf{m}, S)

Step 1a: Put a prior on S

$$(\nu_0 - m - 1) S \sim \text{IW}(\nu_0, (\nu_0 - m - 1) S_0)$$

This thing is called the Inverse-Wishart distribution.

A PDF over SPD matrices!

Being Bayesian: MAP estimates for Gaussians

- v_0 small: "I am not sure about my guess of S_0 "
- v_0 large: "I'm pretty sure about my guess of S_0 "

S_0 : (Roughly) my best guess of S

$$E[S] = S_0$$

Step 1: Put a prior on (\mathbf{m}, S)

Step 1a: Put a prior on S

$$(v_0 - m - 1) S \sim IW(v_0, (v_0 - m - 1) S_0)$$

This thing is called the Inverse-Wishart distribution.

A PDF over SPD matrices!

Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R \sim (\text{i.i.d}) N(\mathbf{m}, S)$
- But you don't know \mathbf{m} or S
- MAP: Which (\mathbf{m}, S) maximizes $p(\mathbf{m}, S | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R)$?

Step 1: Put a prior on (\mathbf{m}, S)

Step 1a: Put a prior on S

$$(v_0 - m - 1) S \sim IW(v_0, (v_0 - m - 1) S_0)$$

Step 1b: Put a prior on $\mathbf{m} | S$

$$\mathbf{m} | S \sim N(\mathbf{m}_0, S / \kappa_0)$$

} Together, " S " and " $\mathbf{m} | S$ " define a joint distribution on (\mathbf{m}, S)

Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R \sim (\text{i.i.d}) N(\mathbf{m}, S)$
- But you don't know \mathbf{m} or S
- MAP: Which (\mathbf{m}, S) maximizes $p(\mathbf{m}, S | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R)$?

\mathbf{m}_0 : My best guess of \mathbf{m} (\mathbf{m}, S)
 $E[\mathbf{m}] = \mathbf{m}_0$

κ_0 small: "I am not sure about my guess of \mathbf{m}_0 "

κ_0 large: "I'm pretty sure about my guess of \mathbf{m}_0 "

$$(v_0 - m - 1)S \sim \text{IW}(v_0, (v_0 - m - 1)S_0)$$

Step 1b: Put a prior on $\mathbf{m} | S$

$$\mathbf{m} | S \sim N(\mathbf{m}_0, S / \kappa_0)$$

Together, " S " and " $\mathbf{m} | S$ " define a joint distribution on (\mathbf{m}, S)

Notice how we are forced to express our ignorance of \mathbf{m} proportionally to S

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 41

Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R \sim (\text{i.i.d}) N(\mathbf{m}, S)$
- But you don't know \mathbf{m} or S
- MAP: Which (\mathbf{m}, S) maximizes $p(\mathbf{m}, S | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R)$?

Step 1: Put a prior on (\mathbf{m}, S)

Why do we use this form of prior?

Step 1a: Put a prior on S

$$(v_0 - m - 1)S \sim \text{IW}(v_0, (v_0 - m - 1)S_0)$$

Step 1b: Put a prior on $\mathbf{m} | S$

$$\mathbf{m} | S \sim N(\mathbf{m}_0, S / \kappa_0)$$

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 42

Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R \sim (\text{i.i.d}) N(\mathbf{m}, S)$
- But you don't know \mathbf{m} or S
- MAP: Which (\mathbf{m}, S) maximizes $p(\mathbf{m}, S | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R)$?

Step 1: Put a prior on (\mathbf{m}, S)

Why do we use this form of prior?

Step 1a: Put a prior on S

Actually, we don't have to

$$(v_0 - m - 1)S \sim \text{IW}(v_0, (v_0 - m - 1)S_0)$$

But it is computationally and algebraically convenient...

Step 1b: Put a prior on $\mathbf{m} | S$

...it's a *conjugate prior*.

$$\mathbf{m} | S \sim N(\mathbf{m}_0, S / \kappa_0)$$

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 43

Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R \sim (\text{i.i.d}) N(\mathbf{m}, S)$
- MAP: Which (\mathbf{m}, S) maximizes $p(\mathbf{m}, S | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R)$?

Step 1: Prior: $(v_0 - m - 1)S \sim \text{IW}(v_0, (v_0 - m - 1)S_0)$, $\mathbf{m} | S \sim N(\mathbf{m}_0, S / \kappa_0)$

Step 2:

$$\bar{\mathbf{x}} = \frac{1}{R} \sum_{k=1}^R \mathbf{x}_k \quad \boldsymbol{\mu}_R = \frac{\mathbf{k}_0 \boldsymbol{\mu}_0 + R \bar{\mathbf{x}}}{\mathbf{k}_0 + R} \quad \mathbf{n}_R = \mathbf{n}_0 + R$$

$$\mathbf{k}_R = \mathbf{k}_0 + R$$

$$(\mathbf{n}_R + m - 1)S_R = (\mathbf{n}_0 + m - 1)S_0 + \sum_{k=1}^R (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T + \frac{(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T}{1/\mathbf{k}_0 + 1/R}$$

Step 3: Posterior: $(v_R + m - 1)S \sim \text{IW}(v_R, (v_R + m - 1)S_R)$,

$$\mathbf{m} | S \sim N(\mathbf{m}_R, S / \kappa_R)$$

Result: $\mathbf{m}^{\text{map}} = \mathbf{m}_R$, $E[S | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R] = S_R$

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 44

Being Bayesian:

- Suppose you have
- MAP: Which (\mathbf{m}, \mathbf{S})

Step 1: Prior: $(\nu_0, \mathbf{m}_0, \mathbf{S}_0)$

Step 2:

$$\bar{\mathbf{x}} = \frac{1}{R} \sum_{k=1}^R \mathbf{x}_k$$

$$\boldsymbol{\mu}_R = \frac{\mathbf{k}_0 \boldsymbol{\mu}_0 + R \bar{\mathbf{x}}}{\mathbf{k}_0 + R}$$

$$\mathbf{n}_R = \mathbf{n}_0 + R$$

$$\mathbf{k}_R = \mathbf{k}_0 + R$$

- Look carefully at what these formulae are doing. It's all very sensible.
- Conjugate priors mean prior form and posterior form are same and characterized by "sufficient statistics" of the data.
- The marginal distribution on \mathbf{m} is a student-t
- One point of view: it's pretty academic if $R > 30$

$$(\mathbf{n}_R + m - 1) \mathbf{S}_R = (\mathbf{n}_0 + m - 1) \mathbf{S}_0 + \sum_{k=1}^R (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T + \frac{(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T}{1/\mathbf{k}_0 + 1/R}$$

Step 3: Posterior: $(\nu_R, \mathbf{m}_R, \mathbf{S}_R) \sim \text{IW}(\nu_R, (\nu_R + m - 1) \mathbf{S}_R)$,

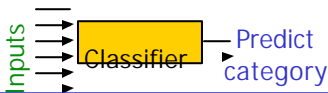
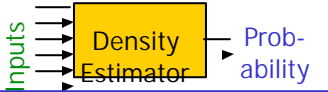
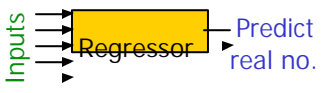
$$\mathbf{m} | \mathbf{S} \sim \text{N}(\mathbf{m}_R, \mathbf{S} / \kappa_R)$$

Result: $\mathbf{m}^{\text{map}} = \mathbf{m}_R$, $E[\mathbf{S} | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_R] = \mathbf{S}_R$

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 45

Where we're at

	Categorical inputs only	Real-valued inputs only	Mixed Real / Cat okay
	Joint BC Naïve BC		Dec Tree
	Joint DE Naïve DE	Gauss DE	
			

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 46

What you should know

- The Recipe for MLE
- What do we sometimes prefer MLE to MAP?
- Understand MLE estimation of Gaussian parameters
- Understand “biased estimator” versus “unbiased estimator”
- Appreciate the outline behind Bayesian estimation of Gaussian parameters

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 47

Useful exercise

- We'd already done some MLE in this class without even telling you!
- Suppose categorical arity- n inputs x_1, x_2, \dots, x_R (i.i.d.) from a multinomial

$$M(p_1, p_2, \dots, p_n)$$

where

$$P(x_k=j|\mathbf{p})=p_j$$

- What is the MLE $\mathbf{p}=(p_1, p_2, \dots, p_n)$?

Copyright © 2001, Andrew W. Moore

Maximum Likelihood: Slide 48