Learning with Maximum Likelihood

Andrew W. Moore
Professor
School of Computer Science
Carnegie Mellon University
www.cs.cmu.edu/~awm
awm@cs.cmu.edu
412-268-7599

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.

 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

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 $x_1, x_2, \ldots x_R \sim$ (i.i.d) $N(\mu,\sigma^2)$
- But you don't know μ

(you do know σ^2)

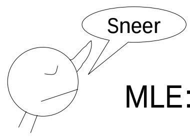
MLE: For which μ is $x_1, x_2, \ldots x_R$ most likely?

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

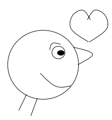
Learning Gaussians from Data

- Suppose you have $x_1, x_2, \ldots x_R$ ~(i.i.d) $N(\mu,\sigma^2)$
- But you don't know μ

(you do know σ^2)



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

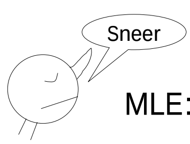


MAP: Which μ maximizes $p(\mu|x_1, x_2, \ldots x_R, \sigma^2)$?

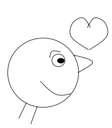
Learning Gaussians from Data

- Suppose you have $x_1, x_2, \ldots x_R$ ~(i.i.d) $N(\mu,\sigma^2)$
- But you don't know μ

(you do know σ^2)



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



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

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

MLE for univariate Gaussian

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

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

Algebra Euphoria

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

= (by i.i.d)

= (monotonicity of log)

= (plug in formula for Gaussian)

= (after simplification)

Algebra Euphoria

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

$$= \arg\max_{\mu} \prod_{i=1}^{R} p(x_i \mid \mu, \sigma^2)$$ (by i.i.d)

$$= \arg\max_{\mu} \sum_{i=1}^{R} \log p(x_i \mid \mu, \sigma^2)$$ (monotonicity of log)

$$= \arg\max_{\mu} \frac{1}{\sqrt{2\pi}\,\sigma} \sum_{i=1}^{R} -\frac{(x_i - \mu)^2}{2\sigma^2}$$ (plug in formula for Gaussian)

$$= \arg\min_{\mu} \sum_{i=1}^{R} (x_i - \mu)^2$$ (after simplification)

Intermission: A General Scalar MLE strategy

Task: Find MLE θ assuming known form for p(Data| θ,stuff)

1. Write LL = log P(Data| θ,stuff)
2. Work out ∂LL/∂θ using high-school calculus
3. Set ∂LL/∂θ=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.

The MLE μ

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

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

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

= (what?)

The MLE μ

$$\mu^{mle} = \arg\max_{\mu} p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2)$$

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

$$= \mu \text{ s.t. } 0 = \frac{\partial \text{LL}}{\partial \mu} = \frac{\partial}{\partial \mu} \sum_{i=1}^{R} (x_i - \mu)^2$$

$$- \sum_{i=1}^{R} 2(x_i - \mu)$$

$$\text{Thus} \quad \mu = \frac{1}{R} \sum_{i=1}^{R} x_i$$

Lawks-a-lawdy!

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

- 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.

A General MLE strategy

Suppose $\theta = (\theta_1, \theta_2, ..., \theta_n)^T$ is a vector of parameters.

Task: Find MLE θ assuming known form for p(Data| θ,stuff)

1. Write LL = log P(Data| θ,stuff)
2. Work out $\partial LL/\partial\theta$ using high-school calculus

$$\frac{\partial LL}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial LL}{\partial \theta_1} \\ \frac{\partial LL}{\partial \theta_2} \\ \vdots \\ \frac{\partial LL}{\partial \theta_n} \end{pmatrix}$$

A General MLE strategy

Suppose θ = (θ_1, θ_2, ..., θ_n)^T is a vector of parameters.

Task: Find MLE θ assuming known form for p(Data| θ,stuff)

1. Write LL = log P(Data| θ,stuff)
2. Work out ∂LL/∂θ using high-school calculus
3. Solve the set of simultaneous equations

$$\frac{\partial LL}{\partial \theta_1} = 0$$
$$\frac{\partial LL}{\partial \theta_2} = 0$$
$$\vdots$$
$$\frac{\partial LL}{\partial \theta_n} = 0$$

A General MLE strategy

Suppose θ = (θ_1, θ_2, ..., θ_n)^T is a vector of parameters.

Task: Find MLE θ assuming known form for p(Data| θ,stuff)

1. Write LL = log P(Data| θ,stuff)
2. Work out ∂LL/∂θ using high-school calculus
3. Solve the set of simultaneous equations

$$\frac{\partial LL}{\partial \theta_1} = 0$$
$$\frac{\partial LL}{\partial \theta_2} = 0$$
$$\vdots$$
$$\frac{\partial LL}{\partial \theta_n} = 0$$

4. Check that you're at a maximum

A General MLE strategy

Suppose θ = (θ_1, θ_2, ..., θ_n)^T is a vector of parameters.

Task: Find MLE θ assuming known form for p(Data| θ,stuff)

1. Write LL = log P(Data| θ,stuff)
2. Work out ∂LL/∂θ using high-school calculus
3. Solve the set of simultaneous equations

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

$$\frac{\partial LL}{\partial \theta_1} = 0$$
$$\frac{\partial LL}{\partial \theta_2} = 0$$
$$\vdots$$
$$\frac{\partial LL}{\partial \theta_n} = 0$$

4. Check that you're at a maximum

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2)
- But you don't know μ or σ^2
- MLE: For which θ =(μ,σ^2) is $x_1, x_2, \ldots x_R$ most likely?

$$\log p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2) = -R(\log \pi + \frac{1}{2}\log \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{R}(x_i - \mu)^2$$

$$\frac{\partial LL}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{R}(x_i - \mu)$$

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

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2)
- But you don't know μ or σ^2
- MLE: For which θ =(μ,σ^2) is $x_1, x_2, \ldots x_R$ most likely?

$$\log p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2) = -R(\log \pi + \frac{1}{2}\log \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{R}(x_i - \mu)^2$$

$$0 = \frac{1}{\sigma^2}\sum_{i=1}^{R}(x_i - \mu)$$

$$0 = -\frac{R}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{R}(x_i - \mu)^2$$

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2)
- But you don't know μ or σ^2
- MLE: For which θ =(μ,σ^2) is $x_1, x_2, \ldots x_R$ most likely?

$$\log p(x_1, x_2, \ldots x_R \mid \mu, \sigma^2) = -R(\log \pi + \frac{1}{2}\log \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{R}(x_i - \mu)^2$$

$$0 = \frac{1}{\sigma^2}\sum_{i=1}^{R}(x_i - \mu) \Rightarrow \mu = \frac{1}{R}\sum_{i=1}^{R}x_i$$

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

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2)
- But you don't know μ or σ^2
- MLE: For which $\theta = (\mu,\sigma^2)$ is $x_1, x_2, \ldots x_R$ most likely?

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

$$\sigma^2_{mle} = \frac{1}{R}\sum_{i=1}^{R} (x_i - \mu^{mle})^2$$

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* $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2) then

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

μ^{mle} is unbiased

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* $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2) then

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

σ^2_{mle} is biased

MLE Variance Bias

- *If* $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2) then

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

Intuition check: consider the case of R=1

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

How could you prove that?

Unbiased estimate of Variance

- *If* $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2) then

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

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

Unbiased estimate of Variance

- *If* $x_1, x_2, \ldots x_R$ ~(i.i.d) N(μ,σ^2) then

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

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

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

Unbiaseditude discussion

- *Which is best?*

$$\sigma^2_{mle} = \frac{1}{R}\sum_{i=1}^{R}(x_i - \mu^{mle})^2$$

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

Answer:

•It depends on the task

•And doesn't make much difference once R--> large

Don't get too excited about being unbiased

- *Assume $x_1, x_2, \ldots x_R$* ~(i.i.d) N(μ,σ^2)
- Suppose we had these estimators for the mean

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

$$\mu^{crap} = x_1$$

Are either of these unbiased?

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

Which is more useful?

MLE for m-dimensional Gaussian

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ ~(i.i.d) N(μ,Σ)
- But you don't know μ or Σ
- MLE: For which θ =(μ,Σ) is $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ most likely?

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

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

MLE for m-dimensional Gaussian

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ ~(i.i.d) N(μ,Σ)
- But you don't know μ or Σ
- MLE: For which θ =(μ,Σ) is $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ most likely?

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

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

Where $1 \le i \le m$

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

And μ_i^{mle} is the i^{th} component of $\boldsymbol{\mu}^{mle}$

MLE for m-dimensional Gaussian

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ ~(i.i.d) N(μ,Σ)
- But you don't know μ or Σ
- MLE: For which θ =(μ,Σ) is $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ most likely?

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

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

Where 1 ≤ i ≤ m, 1 ≤ j ≤ 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 $(i,j)^{th}$ component of Σ^{mle}

$$\sigma_{ij}^{mle} = \frac{1}{R}\sum_{k=1}^{R} \left(\mathbf{x}_{ki} - \mu_i^{mle}\right)\left(\mathbf{x}_{kj} - \mu_j^{mle}\right)$$

MLE for m-dimensional Gaussian

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots$
- But you don't know μ or Σ
- MLE: For which θ =(μ,Σ) is

Q: How would you prove this?

A: Just plug through the MLE recipe.

Note how Σ^{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?

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

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

$$\boldsymbol{\Sigma}^{\text{unbiased}} = \frac{\boldsymbol{\Sigma}^{mle}}{1-\frac{1}{R}} = \frac{1}{R-1}\sum_{k=1}^{R} \left(\mathbf{x}_k - \mu^{mle}\right)\left(\mathbf{x}_k - \mu^{mle}\right)^T$$

Confidence intervals

We need to talk

We need to discuss how accurate we expect μ^{mle} and Σ^{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?

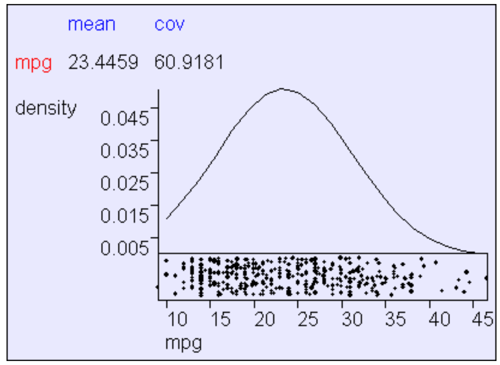
How can we tell? *

How can we survive? *

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

Gaussian MLE in action

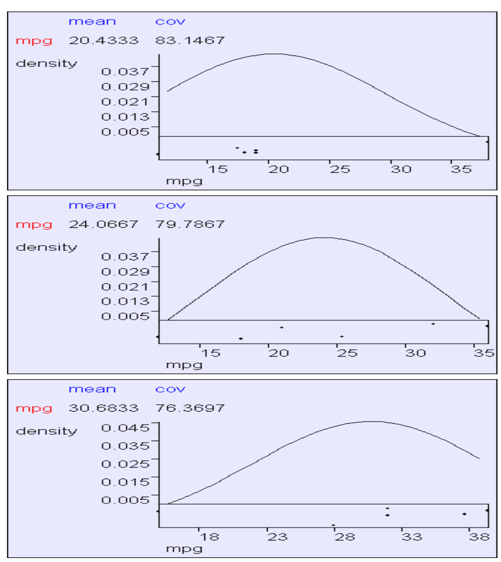
Using R=392 cars from the "MPG" UCI dataset supplied by Ross Quinlan



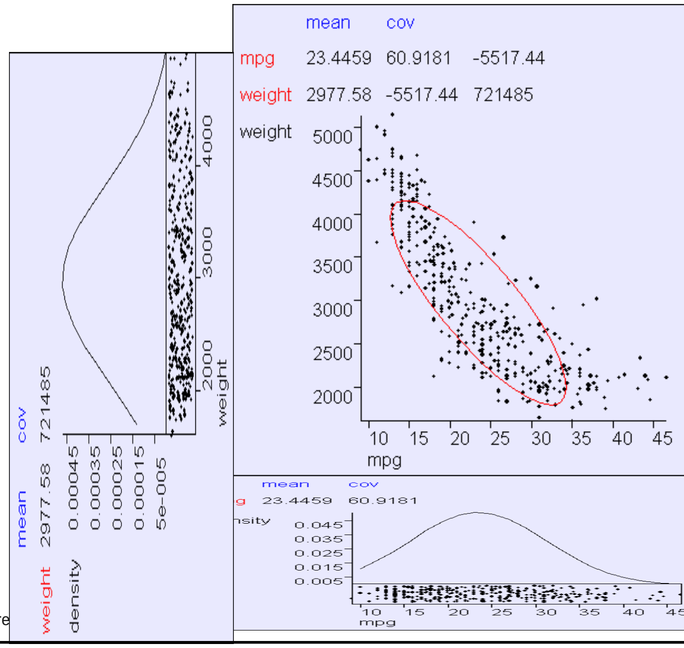
Data-starved Gaussian MLE

Using three subsets of MPG.

Each subset has 6 randomly-chosen cars.



Bivariate MLE in action



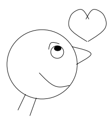
Multivariate MLE

	mean	cov						
mpg	23.4459	60.9181	-10.3529	-657.585	-233.858	-5517.44	9.11551	16.6915
cylinders	5.47194	-10.3529	2.9097	169.722	55.3482	1300.42	-2.37505	-2.17193
displacement	194.412	-657.585	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	-967.228
acceleration	15.5413	9.11551	-2.37505	-156.994	-73.187	-976.815	7.61133	2.95046
modelyear	75.9796	16.6915	-2.17193	-142.572	-59.0364	-967.228	2.95046	13.5699

Covariance matrices are not exciting to look at

Being Bayesian: MAP estimates for Gaussians

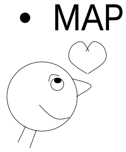
- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ ~(i.i.d) N(μ,Σ)
- But you don't know μ or Σ
- MAP: Which (μ,Σ) maximizes $p(\mu,\Sigma \mid \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R)$?



Step 1: Put a prior on (μ,Σ)

Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ ~(i.i.d) N(μ,Σ)
- But you don't know μ or Σ
- MAP: Which (μ,Σ) maximizes $p(\mu,\Sigma \mid \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R)$?



Step 1: Put a prior on (μ,Σ)

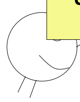
Step 1a: Put a prior on Σ

$(\nu_0\text{-m-1})\ \Sigma \sim IW(\nu_0, (\nu_0\text{-m-1})\ \Sigma_0)$

This thing is called the Inverse-Wishart distribution.

A PDF over SPD matrices!

Being Bayesian: MAP estimates for Gaussians



ν_0 small: "I am not sure about my guess of Σ_0"

ν_0 large: "I'm pretty sure about my guess of Σ_0"

Σ_0 : (Roughly) my best guess of Σ

$E[\Sigma] = \Sigma_0$

Step 1: Put a prior on (μ,Σ)

Step 1a: Put a prior on Σ

$(\nu_0\text{-m-1})\,\Sigma \sim \text{IW}(\nu_0, (\nu_0\text{-m-1})\,\Sigma_0)$

This thing is called the Inverse-Wishart distribution.

A PDF over SPD matrices!

Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R \sim$(i.i.d) $N(\mu,\Sigma)$
- But you don't know μ or Σ
- MAP: Which (μ,Σ) maximizes $p(\mu,\Sigma \mid \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R)$?

Step 1: Put a prior on (μ,Σ)

Step 1a: Put a prior on Σ

$(\nu_0\text{-m-1})\Sigma \sim \text{IW}(\nu_0, (\nu_0\text{-m-1})\Sigma_0)$

Step 1b: Put a prior on $\mu \mid \Sigma$

$\mu \mid \Sigma \sim N(\mu_0, \Sigma / \kappa_0)$

Together, "Σ" and "$\mu \mid \Sigma$" define a joint distribution on (μ,Σ)

Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ ~(i.i.d) N(μ,Σ)
- But you don't know μ or Σ
- MAP: Which (μ,Σ) maximi...)?

μ 0 : My best guess of μ
E[μ] = μ 0

κ_0 small: "I am not sure about my guess of μ 0 "

κ_0 large: "I'm pretty sure about my guess of μ 0 "

... μ,Σ)

... Σ

$(\nu_0\text{-m-1})\Sigma$ ~ ... ν_0, $(\nu_0\text{-m-1})\Sigma$ ...

Step 1b: Put a pr...r on μ | Σ

$\mu \mid \Sigma \sim N(\mu_0 , \Sigma / \kappa_0)$

Together, "Σ" and "μ | Σ" define a joint distribution on (μ,Σ)

Notice how we are forced to express our ignorance of μ proportionally to Σ

Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ ~(i.i.d) N(μ,Σ)
- But you don't know μ or Σ
- MAP: Which (μ,Σ) maximizes p(μ,Σ |$\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$)?

Step 1: Put a prior on (μ,Σ)

Step 1a: Put a prior on Σ

$(\nu_0\text{-m-1})\Sigma \sim IW(\nu_0, (\nu_0\text{-m-1})\Sigma_0)$

Step 1b: Put a prior on μ | Σ

$\mu \mid \Sigma \sim N(\mu_0 , \Sigma / \kappa_0)$

Why do we use this form of prior?

Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ ~(i.i.d) N(μ,Σ)
- But you don't know μ or Σ
- MAP: Which (μ,Σ) maximizes p(μ,Σ | $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$)?

Step 1: Put a prior on (μ,Σ)

Step 1a: Put a prior on Σ

$(\nu_0\text{-m-1})\Sigma \sim \text{IW}(\nu_0, (\nu_0\text{-m-1})\Sigma_0)$

Step 1b: Put a prior on μ | Σ

$\mu \mid \Sigma \sim N(\mu_0, \Sigma / \kappa_0)$

Why do we use this form of prior?

Actually, we don't have to

But it is computationally and algebraically convenient…

…it's a *conjugate prior.*

Being Bayesian: MAP estimates for Gaussians

- Suppose you have $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$ ~(i.i.d) N(μ,Σ)
- MAP: Which (μ,Σ) maximizes p(μ,Σ | $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R$)?

Step 1: Prior: $(\nu_0\text{-m-1})\,\Sigma \sim \text{IW}(\nu_0, (\nu_0\text{-m-1})\,\Sigma_0)$, $\mu \mid \Sigma \sim N(\mu_0, \Sigma / \kappa_0)$

Step 2:

$$\bar{\mathbf{x}} = \frac{1}{R}\sum_{k=1}^{R}\mathbf{x}_k \qquad \boldsymbol{\mu}_R = \frac{\kappa_0\boldsymbol{\mu}_0 + R\bar{\mathbf{x}}}{\kappa_0 + R} \qquad \nu_R = \nu_0 + R \qquad \kappa_R = \kappa_0 + R$$

$$(\nu_R + m - 1)\boldsymbol{\Sigma}_R = (\nu_0 + m - 1)\boldsymbol{\Sigma}_0 + \sum_{k=1}^{R}(\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T + \frac{(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T}{1/\kappa_0 + 1/R}$$

Step 3: Posterior: $(\nu_R\text{+m-1})\Sigma \sim \text{IW}(\nu_R, (\nu_R\text{+m-1})\,\Sigma_R)$,

$\mu \mid \Sigma \sim N(\mu_R, \Sigma / \kappa_R)$

Result: $\mu^{map} = \mu_R$, $E[\Sigma \mid \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R] = \Sigma_R$

Being Bayesian:

- Suppose you hav
- MAP: Which (μ,Σ)

•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 μ is a student-t
•One point of view: it's pretty academic if R > 30

Step 1: Prior: $(\nu_0\text{-m-1})\ \Sigma$ ~

Step 2:

$$\bar{\mathbf{x}} = \frac{1}{R}\sum_{k=1}^{R} \mathbf{x}_k \qquad \boldsymbol{\mu}_R = \frac{\kappa_0 \boldsymbol{\mu}_0 + R\bar{\mathbf{x}}}{\kappa_0 + R} \qquad \nu_R = \nu_0 + R \qquad \kappa_R = \kappa_0 + R$$

$$(\nu_R + m - 1)\boldsymbol{\Sigma}_R = (\nu_0 + m - 1)\boldsymbol{\Sigma}_0 + \sum_{k=1}^{R}(\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T + \frac{(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^T}{1/\kappa_0 + 1/R}$$

Step 3: Posterior: $(\nu_R\text{+m-1})\Sigma \sim \text{IW}(\nu_R, (\nu_R\text{+m-1})\ \Sigma_R)$,

$\mu \mid \Sigma \sim N(\mu_R , \Sigma / \kappa_R)$

Result: $\mu^{map} = \mu_R$, $E[\Sigma \mid \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R] = \Sigma_R$

Where we're at

	Categorical inputs only	Real-valued inputs only	Mixed Real / Cat okay
Inputs → Classifier → Predict category	Joint BC Naïve BC		Dec Tree
Inputs → Density Estimator → Prob-ability	Joint DE Naïve DE	Gauss DE	
Inputs → Regressor → Predict real no.			

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

Useful exercise

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

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

where

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

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

