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 $x_1$, $x_2$, … $x_R \sim$ (i.i.d) $N(\mu, \sigma^2)$
• But you don’t know $\mu$
  (you do know $\sigma^2$)

  MLE: For which $\mu$ is $x_1$, $x_2$, … $x_R$ most likely?

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

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

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

MAP: Which $\mu$ 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 $x_1, x_2, \ldots, x_R \sim \text{(i.i.d) } N(\mu, \sigma^2)$
- But you don’t know $\mu$ (you do know $\sigma^2$)
- MLE: For which $\mu$ 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

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

$$= \quad \text{(by i.i.d)}$$

$$= \quad \text{(monotonicity of log)}$$

$$= \quad \text{(plug in formula for Gaussian)}$$

$$= \quad \text{(after simplification)}$$
Intermission: A General Scalar
MLE strategy

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

1. Write $LL = \log p(\text{Data} \mid \theta, \text{stuff})$
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 $\theta$
4. Solve it*
5. Check that you’ve found a maximum rather than a minimum or saddle-point, and be careful if $\theta$ 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 $\mu$

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

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

$= \mu \quad \text{s.t.} \quad 0 = \frac{\partial LL}{\partial \mu} = \quad \text{(what?)}$
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 \( \theta \) 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

\[
\frac{\partial LL}{\partial \theta} = \begin{pmatrix}
\frac{\partial LL}{\partial \theta_1} \\
\frac{\partial LL}{\partial \theta_2} \\
\vdots \\
\frac{\partial LL}{\partial \theta_n}
\end{pmatrix}
\]
A General MLE strategy

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

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

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

\[
\begin{align*}
\frac{\partial LL}{\partial \theta_1} &= 0 \\
\frac{\partial LL}{\partial \theta_2} &= 0 \\
&\vdots \\
\frac{\partial LL}{\partial \theta_n} &= 0
\end{align*}
\]
A General MLE strategy

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

Task: Find MLE $\theta$ 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. Solve the set of simultaneous equations

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

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

4. Check that you're at a maximum

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

MLE for univariate Gaussian

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

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

$$\begin{align*}
\frac{\partial LL}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^{R} (x_i - \mu) \\
\frac{\partial LL}{\partial \sigma^2} &= -\frac{R}{2\sigma^4} + \frac{1}{2\sigma^4} \sum_{i=1}^{R} (x_i - \mu)^2
\end{align*}$$
MLE for univariate Gaussian

• Suppose you have \( x_1, x_2, \ldots, x_R \sim \text{(i.i.d) } N(\mu, \sigma^2) \)

• But you don’t know \( \mu \) or \( \sigma^2 \)

• MLE: For which \( \theta = (\mu, \sigma^2) \) 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 \( x_1, x_2, \ldots, x_R \sim \text{(i.i.d) } N(\mu, \sigma^2) \)
- But you don’t know \( \mu \) or \( \sigma^2 \)
- MLE: For which \( \theta = (\mu, \sigma^2) \) is \( x_1, x_2, \ldots, x_R \) most likely?

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

\[
\sigma_{mle}^2 = \frac{1}{R} \sum_{i=1}^{R} (x_i - \mu^{mle})^2
\]

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

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

\( \mu^{mle} \) 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 \( x_1, x_2, \ldots, x_R \sim \text{(i.i.d) } N(\mu, \sigma^2) \) then

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

\( \hat{\sigma}^2_{mle} \) is biased

MLE Variance Bias

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

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

Intuition check: consider the case of \( R=1 \)

Why should our guts expect that \( \sigma^2_{mle} \) would be an underestimate of true \( \sigma^2 \)?

How could you prove that?
Unbiased estimate of Variance

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

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

So define

$$\sigma^2_{\text{unbiased}} = \frac{\sigma^2_{\text{mle}}}{\left(1 - \frac{1}{R}\right)}$$

So $E[\sigma^2_{\text{unbiased}}] = \sigma^2$

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

• Which is best?

\[
\sigma_{mle}^2 = \frac{1}{R} \sum_{i=1}^{R} (x_i - \mu_{mle})^2
\]

\[
\sigma_{unbiased}^2 = \frac{1}{R-1} \sum_{i=1}^{R} (x_i - \mu_{mle})^2
\]

Answer:
• It depends on the task
• And doesn't make much difference once R --> large

Don’t get too excited about being unbiased

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

\[
\mu_{\text{suboptimal}} = \frac{1}{R + 7\sqrt{R}} \sum_{i=1}^{R} x_i
\]

\[
\mu_{\text{crap}} = x_i
\]

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 \( x_1, x_2, \ldots, x_R \sim \text{(i.i.d)} \, N(\mu, \Sigma) \)
- But you don’t know \( \mu \) or \( \Sigma \)
- MLE: For which \( \theta = (\mu, \Sigma) \) is \( x_1, x_2, \ldots, x_R \) most likely?

\[
\mu^{mle} = \frac{1}{R} \sum_{k=1}^{R} x_k
\]

\[
S^{mle} = \frac{1}{R} \sum_{k=1}^{R} (x_k - \mu^{mle})(x_k - \mu^{mle})^T
\]

\[
\mu_{i}^{mle} = \frac{1}{R} \sum_{k=1}^{R} x_{ki}
\]

Where \( 1 \leq i \leq m \)

And \( x_{ki} \) is value of the \( i^{th} \) component of \( x_k \)

(\( i^{th} \) attribute of the \( k^{th} \) record)

And \( \mu_{i}^{mle} \) is the \( i^{th} \) component of \( \mu^{mle} \)
MLE for m-dimensional Gaussian

• Suppose you have \( x_1, x_2, \ldots, x_R \sim \text{i.i.d.} \ N(\mu, \Sigma) \)
• But you don’t know \( \mu \) or \( \Sigma \)
• MLE: For which \( \theta = (\mu, \Sigma) \) is \( x_1, x_2, \ldots, x_R \) most likely?

\[
\mu_{\text{mle}} = \frac{1}{R} \sum_{k=1}^{R} x_k
\]

\[
S_{\text{mle}} = \frac{1}{R} \sum_{k=1}^{R} (x_k - \mu_{\text{mle}})(x_k - \mu_{\text{mle}})^T
\]

\[
\sigma_{ij_{\text{mle}}} = \frac{1}{R} \sum_{k=1}^{R} (x_{ki} - \mu_{ij_{\text{mle}}})(x_{kj} - \mu_{ij_{\text{mle}}})
\]

Where \( 1 \leq i \leq m, 1 \leq j \leq m \)

And \( x_{ki} \) is value of the \( i \)th component of \( x_k \) (the \( i \)th attribute of the \( k \)th record)

And \( \sigma_{ij_{\text{mle}}} \) is the \( (i,j) \)th component of \( \Sigma_{\text{mle}} \)

Q: How would you prove this?
A: Just plug through the MLE recipe.

Note how \( \Sigma_{\text{mle}} \) 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?
Confidence intervals

We need to talk

We need to discuss how accurate we expect $\mu^{\text{mle}}$ and $\Sigma^{\text{mle}}$ 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

<table>
<thead>
<tr>
<th>mean</th>
<th>cov</th>
</tr>
</thead>
<tbody>
<tr>
<td>mpg</td>
<td>23.4459  80.9181  -10.3529  -857.585  -233.858  -5517.44  9.11551  18.8815</td>
</tr>
<tr>
<td>cylinders</td>
<td>5.47194  -10.3629  2.0097  180.722  55.3482  1200.42  -2.37606  -2.17133</td>
</tr>
<tr>
<td>displacement</td>
<td>194.412  -557.685  169.722  10950.4  3614.03  82929.1  -156.994  -142.572</td>
</tr>
<tr>
<td>horsepower</td>
<td>104.469  -233.858  55.3482  3614.03  1481.57  28285.6  -73.187  -59.0364</td>
</tr>
<tr>
<td>weight</td>
<td>2977.58  -5517.44  1300.42  82929.1  28285.6  721485  -978.815  -987.228</td>
</tr>
<tr>
<td>modelyear</td>
<td>75.9796  16.6915  -2.17193  -142.572  -59.0364  -967.228  2.95046  13.5698</td>
</tr>
</tbody>
</table>

Covariance matrices are not exciting to look at.
Being Bayesian: MAP estimates for Gaussians

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

  Step 1: Put a prior on $(\mu, \Sigma)$

Step 1a: Put a prior on $\Sigma$

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

  This thing is called the Inverse-Wishart distribution.

  A PDF over SPD matrices!
Being Bayesian: MAP estimates for Gaussians

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

Step 1: Put a prior on \( (\mu, \Sigma) \)
Step 1a: Put a prior on \( \Sigma \)
\[
(\nu_0 - m - 1) \Sigma \sim \text{IW}(\nu_0, (\nu_0 - m - 1) \Sigma_0)
\]
This thing is called the Inverse-Wishart distribution.
A PDF over SPD matrices!

\( \nu_0 \) small: “I am not sure about my guess of \( \Sigma_0 \)”
\( \nu_0 \) large: “I’m pretty sure about my guess of \( \Sigma_0 \)”

\( \Sigma_0 \): (Roughly) my best guess of \( \Sigma \)
\[
E[\Sigma] = \Sigma_0
\]
Being Bayesian: MAP estimates for Gaussians

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

\[ \mu_0 : \text{My best guess of } \mu \]
\[ E[\mu] = \mu_0 \]

\[ (\nu_0-m-1)\Sigma \sim IW(\nu_0, (\nu_0-m-1)\Sigma_0) \]

Step 1b: Put a prior on \( \mu | \Sigma \)
\[ \mu | \Sigma \sim N(\mu_0, \Sigma / \kappa_0) \]

\( \kappa_0 \) small: “I am not sure about my guess of \( \mu_0 \)”

\( \kappa_0 \) large: “I’m pretty sure about my guess of \( \mu_0 \)”

Notice how we are forced to express our ignorance of \( \mu \) proportionally to \( \Sigma \)

Why do we use this form of prior?

Together, “\( \Sigma \)” and “\( \mu | \Sigma \)” define a joint distribution on \((\mu, \Sigma)\)
Being Bayesian: MAP estimates for Gaussians

• Suppose you have \( \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_R \sim \text{i.i.d.} \ N(\mu, \Sigma) \)
• But you don’t know \( \mu \) or \( \Sigma \)
• MAP: Which \( (\mu, \Sigma) \) maximizes \( p(\mu, \Sigma | \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_R) \)?

Step 1: Put a prior on \( (\mu, \Sigma) \)

Step 1a: Put a prior on \( \Sigma \)
\[ (\nu_0 - m - 1) \Sigma \sim IW(\nu_0, (\nu_0 - m - 1) \Sigma_0) \]

Step 1b: Put a prior on \( \mu | \Sigma \)
\[ \mu | \Sigma \sim N(\mu_0, \Sigma / \kappa_0) \]

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.

Result: \( \mu_{\text{map}} = \mu_R, \ E[\Sigma | \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_R] = \Sigma_R \)
Being Bayesian:

- Suppose you have $x_1, x_2, \ldots, x_R \sim \text{i.i.d. } N(\mu, \Sigma)$
- MAP: Which $(\mu, \Sigma)$ maximizes $p(\mu, \Sigma | x_1, x_2, \ldots, x_R)$?

**Step 1:** Prior:

- $\bar{x} = \frac{1}{R} \sum_{k=1}^{R} x_k$
- $\mu_R = \frac{\kappa_0 \mu_0 + R \bar{x}}{\kappa_0 + R}$
- $\kappa_R = \kappa_0 + R$

**Step 2:**

- $(\nu_R + m - 1) \Sigma = (\nu_0 + m - 1) \Sigma_0 + \sum_{k=1}^{R} (x_k - \bar{x})(x_k - \bar{x})^T + \frac{(\bar{x} - \mu_0)(\bar{x} - \mu_0)^T}{1/\kappa_0 + 1/R}$

**Step 3:** Posterior:

- $(\nu_R + m - 1) \Sigma \sim \text{IW}(\nu_R, (\nu_R + m - 1) \Sigma_R)$
- $\mu | \Sigma \sim N(\mu_R, \Sigma_R / \kappa_R)$

Result: $\mu_{\text{map}} = \mu_R$, $E[\Sigma | x_1, x_2, \ldots, x_R] = \Sigma_R$

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 $\mu$ is a student-t.

One point of view: it's pretty academic if $R > 30$

---

**Where we’re at**

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Classifier</th>
<th>Predict</th>
<th>Categorical inputs only</th>
<th>Real-valued inputs only</th>
<th>Mixed Real / Cat okay</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Joint BC</td>
<td>Naive BC</td>
<td>Dec Tree</td>
</tr>
<tr>
<td>Density Estimator</td>
<td>Probability</td>
<td></td>
<td>Joint DE</td>
<td>Naive DE</td>
<td>Gauss DE</td>
</tr>
<tr>
<td>Regressor</td>
<td>Predict</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 $x_1, x_2, \ldots x_R \sim \text{(i.i.d.) from a multinomial} M(p_1, p_2, \ldots p_n)$
  where
  \[ P(x_k = j | \mathbf{p}) = p_j \]
• What is the MLE $\mathbf{p}=(p_1, p_2, \ldots p_n)$?