10-701 Machine Learning Recitation #3

MLE, MAP, and Vector/Matrix Differentiation

Calvin Murdock

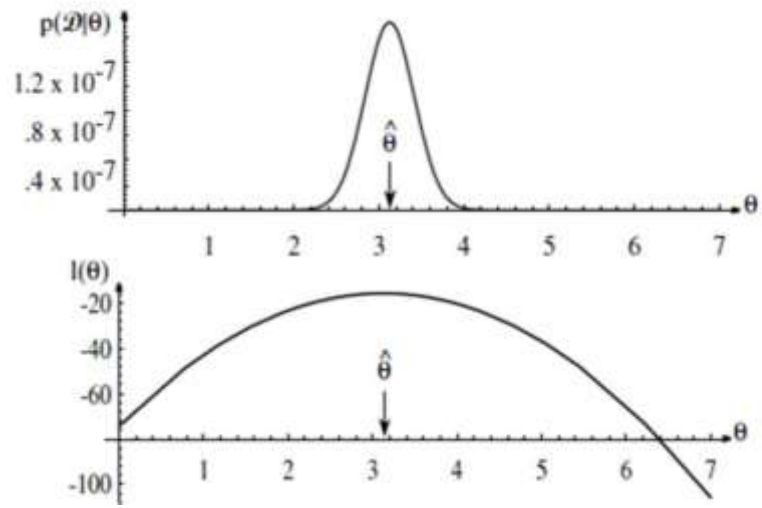
cmurdock@andrew.cmu.edu

(Some slides borrowed from Andrew Moore, Steven Nydick, etc.)

Maximum Likelihood Estimation (MLE)

$$\mathcal{L}(\boldsymbol{\theta}) = \ln p(\mathcal{D}|\boldsymbol{\theta}) = \ln \prod_{i=1}^{N} p\left(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}\right) = \sum_{i=1}^{N} \ln p\left(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}\right)$$

$$\widehat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{N} \ln p\left(\boldsymbol{x}^{(i)}|\boldsymbol{\theta}\right)$$



MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \ldots x_R \sim N(\mu,\sigma^2)$
- 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 $p(\mu | x_1, x_2, \ldots x_R, \sigma^2)$?

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \ldots x_R \sim N(\mu,\sigma^2)$
- 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?

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

MLE for univariate Gaussian

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

MLE for univariate Gaussian

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

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

(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 $\partial LL/\partial\theta$ 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.

The MLE

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

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

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

= (what?)

The MLE

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

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

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

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

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

The MLE

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

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

A General MLE strategy

Suppose $\theta = (\theta_1, \theta_2, \ldots, \theta_n)^T$ 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 the gradient $\partial LL / \partial \theta$ using high-school calculus

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

A General MLE strategy

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

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

1. Write LL = log P(Data| θ,stuff)
2. Work out the gradient $\partial LL/\partial\theta$ 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 $\theta = (\theta_1, \theta_2, \ldots, \theta_n)^T$ 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 the gradient $\partial LL/\partial\theta$ 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 $\theta = (\theta_1, \theta_2, \ldots, \theta_n)^T$ 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 the gradient $\partial LL/\partial\theta$ 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) $N(\mu,\sigma^2)$
- 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?

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

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

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

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) $N(\mu,\sigma^2)$
- 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?

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

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

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

MLE for univariate Gaussian

- Suppose you have $x_1, x_2, \ldots x_R$ ~(i.i.d) $N(\mu,\sigma^2)$
- 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) $N(\mu,\sigma^2)$ 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) $N(\mu,\sigma^2)$ 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) $N(\mu,\sigma^2)$ 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, … 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 $$\sigma^2_{\text{unbiased}} = \frac{\sigma^2_{mle}}{\left(1 - \frac{1}{R}\right)}$$ 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 $$\sigma^2_{\text{unbiased}} = \frac{\sigma^2_{mle}}{\left(1 - \frac{1}{R}\right)}$$ 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) $N(\mu,\sigma^2)$
- 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?

Maximum Conditional Likelihood Estimation (MCLE)

- Same as MLE except with conditional likelihood.
  - E.g. for regression: Given $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, learn $\boldsymbol{\theta}$

Output

Given input

Gaussian noise

$$y = f_{\boldsymbol{\theta}}(\boldsymbol{x}) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Deterministic function with parameters θ

Equivalent to:

mean μ

$$y \sim \mathcal{N}(f_\theta(\boldsymbol{x}), \sigma^2)$$

Maximum Conditional Likelihood Estimation (MCLE)

- Same as MLE except with conditional likelihood.
  - E.g. for regression: Given $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, learn $\boldsymbol{\theta}$

Output

Given input

Gaussian noise

$$y = f_{\boldsymbol{\theta}}(\boldsymbol{x}) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Deterministic function with parameters θ

Equivalent to:

$$y \sim \mathcal{N}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), \sigma^2)$$

Standard MLE: $\mu^{mle} = \arg\max_{\mu} p(\, y_1, \dots, y_n \mid \mu, \sigma^2)$

Maximum Conditional Likelihood Estimation (MCLE)

- Same as MLE except with conditional likelihood.
  - E.g. for regression: Given $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, learn $\boldsymbol{\theta}$

Output

Given input

Gaussian noise

$$y = f_{\boldsymbol{\theta}}(\boldsymbol{x}) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Deterministic function with parameters θ

Equivalent to:

mean $\mu = f_{\boldsymbol{\theta}}(\boldsymbol{x})$

$$y \sim \mathcal{N}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), \sigma^2)$$

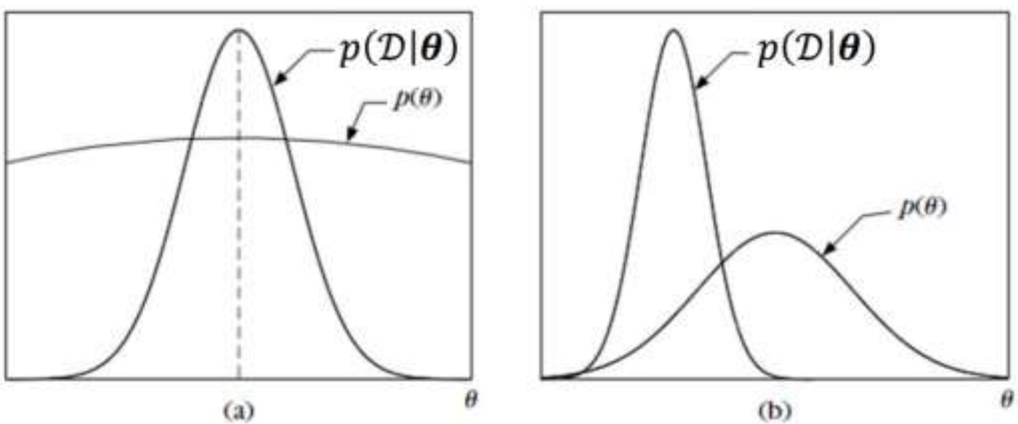
Conditional MLE: $\boldsymbol{\theta}^{mle} = \arg\max_{\boldsymbol{\theta}} p(y_1, \ldots, y_n \mid \boldsymbol{\theta}, \sigma^2, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$

Maximum a Posteriori (MAP) Estimation

$$\widehat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\boldsymbol{\theta}|\mathcal{D})$$

- Since $p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})$

$$\widehat{\boldsymbol{\theta}}_{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$



$\widehat{\boldsymbol{\theta}}_{MAP} \cong \widehat{\boldsymbol{\theta}}_{ML}$ $\qquad$ $\widehat{\boldsymbol{\theta}}_{MAP} > \widehat{\boldsymbol{\theta}}_{ML}$

Being Bayesian: MAP estimates for Gaussians

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

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) $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$
- But you don't know $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$
- MAP: Which $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ maximizes $p(\boldsymbol{\mu},\boldsymbol{\Sigma} \mid \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R)$?

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

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

Step 1: Put a prior on (μ,Σ)

Step 1a: Put a prior on Σ

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

This thing is called the Inverse-Wishart distribution.

A PDF over SPD matrices!

Being Bayesian: MAP estimates for Gaussians

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

Step 1: Put a prior on (μ,Σ)

Step 1a: Put a prior on Σ

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

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) $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$
- But you don't know $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$
- MAP: Which $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ maximizes $p(\boldsymbol{\mu},\boldsymbol{\Sigma} \mid \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R)$?

Step 1: Put a prior on (μ,Σ)

Step 1a: Put a prior on Σ

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

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

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

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) $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$
- But you don't know $\boldsymbol{\mu}$ or $\boldsymbol{\Sigma}$
- MAP: Which $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ maximizes $p(\boldsymbol{\mu},\boldsymbol{\Sigma} \mid \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R)$?

Step 1: Put a prior on (μ,Σ)

Step 1a: Put a prior on Σ

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

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

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

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) $N(\boldsymbol{\mu},\Sigma)$
- MAP: Which $(\boldsymbol{\mu},\Sigma)$ maximizes $p(\boldsymbol{\mu},\Sigma \mid \mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_R)$?

Step 1: Prior: $(\nu_0-m-1)\,\Sigma \sim IW(\nu_0, (\nu_0-m-1)\,\Sigma_0)$, $\boldsymbol{\mu} \mid \Sigma \sim N(\boldsymbol{\mu}_0, \Sigma / \kappa_0)$

Step 2:

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

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

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

$\boldsymbol{\mu} \mid \Sigma \sim N(\boldsymbol{\mu}_R, \Sigma / \kappa_R)$

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

Being Bayesian: M

- Suppose you have **x**
- MAP: Which (μ,Σ) m

•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 $\boldsymbol{\mu}$ 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:

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

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

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

$\boldsymbol{\mu} \mid \Sigma \sim \mathrm{N}(\boldsymbol{\mu}_R, \Sigma / \kappa_R)$

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

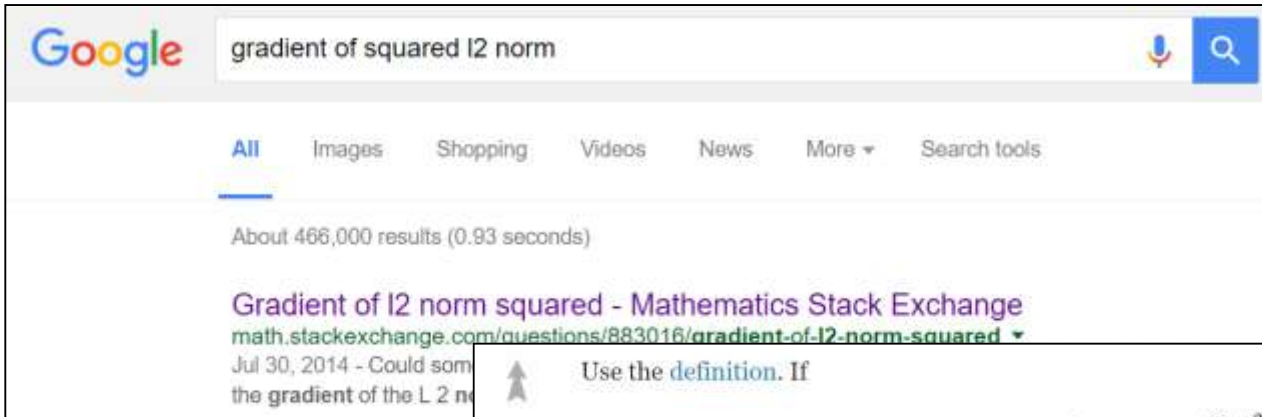
Vector/Matrix Derivatives

- Best reference: **The Matrix Cookbook** (www.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf)

Contents

Vector/Matrix Derivatives

- Second best reference: **Google**



Google

gradient of squared l2 norm

All Images Shopping Videos News More Search tools

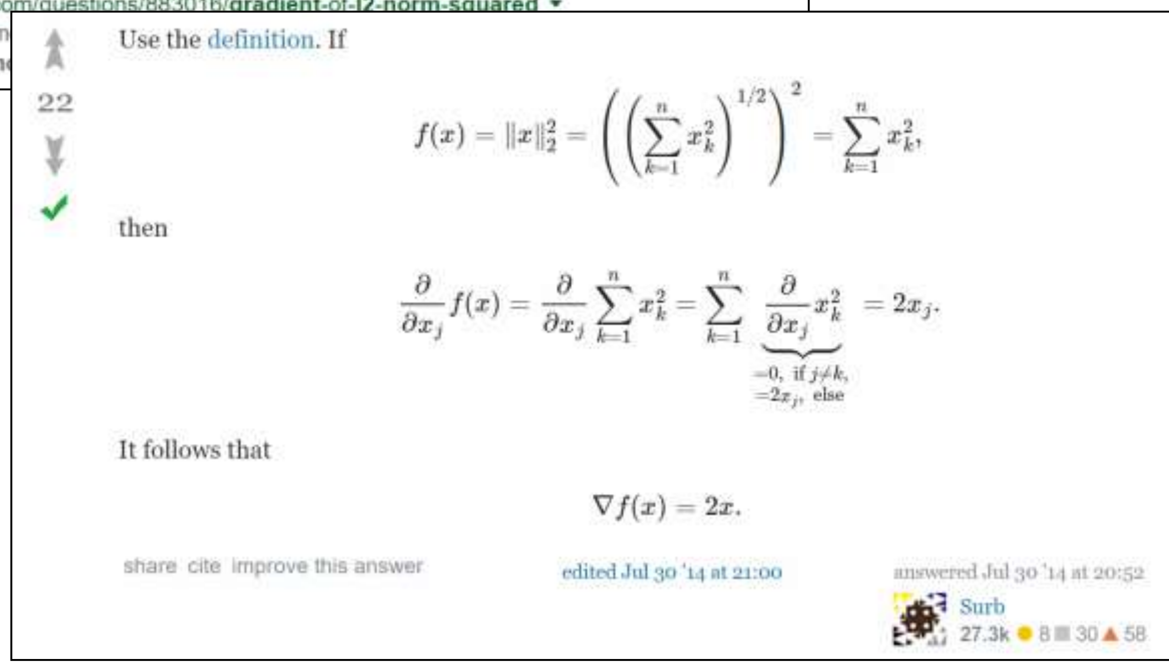
About 466,000 results (0.93 seconds)

Gradient of l2 norm squared - Mathematics Stack Exchange

math.stackexchange.com/questions/883016/gradient-of-l2-norm-squared

Jul 30, 2014 - Could som…

the **gradient** of the L 2 n…



22

Use the definition. If

$$f(x) = \|x\|_2^2 = \left(\left(\sum_{k=1}^{n} x_k^2\right)^{1/2}\right)^2 = \sum_{k=1}^{n} x_k^2,$$

then

$$\frac{\partial}{\partial x_j} f(x) = \frac{\partial}{\partial x_j} \sum_{k=1}^{n} x_k^2 = \sum_{k=1}^{n} \underbrace{\frac{\partial}{\partial x_j} x_k^2}_{\substack{=0, \text{ if } j \neq k, \\ =2x_j, \text{ else}}} = 2x_j.$$

It follows that

$$\nabla f(x) = 2x.$$

share cite improve this answer

edited Jul 30 '14 at 21:00

answered Jul 30 '14 at 20:52

Surb

27.3k ● 8 ■ 30 ▲ 58

Vector Gradient

A Gradient is the derivative of a scalar with respect to a vector.

scalar function

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left(\left[\frac{\partial f(\mathbf{x})}{\partial x_1} \right] \quad \left[\frac{\partial f(\mathbf{x})}{\partial x_2} \right] \quad \cdots \quad \left[\frac{\partial f(\mathbf{x})}{\partial x_n} \right] \right)^T$$

parameter vector

If we have the function: $f(\mathbf{x}) = 2x_1x_2 + x_2^2 + x_1x_3^2$, then the Gradient is

$$\begin{aligned} \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} &= \left(\left[\frac{\partial f(\mathbf{x})}{\partial x_1} \right] \quad \left[\frac{\partial f(\mathbf{x})}{\partial x_2} \right] \quad \left[\frac{\partial f(\mathbf{x})}{\partial x_3} \right] \right)^T \\ &= \left[2x_2 + x_3^2 \quad 2x_1 + 2x_2 \quad 2x_1x_3 \right]^T \end{aligned}$$

Matrix Gradient

scalar function

parameter matrix

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \begin{pmatrix} \left[\frac{\partial f(\mathbf{X})}{\partial x_{1,1}}\right] & \cdots & \left[\frac{\partial f(\mathbf{X})}{\partial x_{1,m}}\right] \\ \vdots & \ddots & \vdots \\ \left[\frac{\partial f(\mathbf{X})}{\partial x_{n,1}}\right] & \cdots & \left[\frac{\partial f(\mathbf{X})}{\partial x_{n,m}}\right] \end{pmatrix}$$

Jacobian

A Jacobian is a the derivative of a vector with respect to a transposed vector.

vector function

parameter vector

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}^T} = \begin{pmatrix} \left[\frac{\partial f_1(\mathbf{x})}{\partial x_1}\right] & \cdots & \left[\frac{\partial f_1(\mathbf{x})}{\partial x_n}\right] \\ \vdots & \cdots & \vdots \\ \left[\frac{\partial f_k(\mathbf{x})}{\partial x_1}\right] & \cdots & \left[\frac{\partial f_k(\mathbf{x})}{\partial x_n}\right] \end{pmatrix} = \begin{pmatrix} \left[\frac{\partial f_1(\mathbf{x})}{\partial \mathbf{x}}\right]^T \\ \vdots \\ \underbrace{\left[\frac{\partial f_k(\mathbf{x})}{\partial \mathbf{x}}\right]^T}_{\nabla f_k} \end{pmatrix}$$

If we have the function

$$\mathbf{f}(\mathbf{x}) = \left[3x_1^2 + x_2 \quad \ln(x_1) \quad \sin(x_2)\right]^T$$

Then the Jacobian is

$$\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}^T} = \begin{pmatrix} 6x_1 & 1 \\ \frac{1}{x_1} & 0 \\ 0 & \cos(x_2) \end{pmatrix}$$

Hessian

The Hessian is derivative of a Gradient with respect to a transposed vector.

scalar function

parameter vector

$$\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} = \begin{pmatrix} \left[\frac{\partial f(\mathbf{x})}{\partial x_1^2}\right] & \cdots & \left[\frac{\partial f(\mathbf{x})}{\partial x_1 \partial x_n}\right] \\ \vdots & \ddots & \vdots \\ \left[\frac{\partial f(\mathbf{x})}{\partial x_n \partial x_1}\right] & \cdots & \left[\frac{\partial f(\mathbf{x})}{\partial x_n^2}\right] \end{pmatrix}$$

Because our above Gradient is

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} 2x_2 + x_3^2 & 2x_1 + 2x_2 & 2x_1 x_3 \end{bmatrix}^T$$

The Hessian would be

$$\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} = \begin{pmatrix} 0 & 2 & 2x_3 \\ 2 & 2 & 0 \\ 2x_3 & 0 & 2x_1 \end{pmatrix}$$

There are two ways of computing gradients…

$$f(\boldsymbol{\theta}) = \|\boldsymbol{x} - \boldsymbol{\theta}\|^{\mathbf{2}} \qquad \frac{df}{d\boldsymbol{\theta}} = ?$$

1. Compute each element of gradient using scalar partial derivatives:

$$f(\boldsymbol{\theta}) = \sum_j \left(x_j - \theta_j\right)^2 \qquad \left(\frac{df}{d\boldsymbol{\theta}}\right)_j = \frac{\partial f}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}\left(x_j - \theta_j\right)^2 = 2(\theta_j - x_j)$$

(using chain rule, power rule)

2. Directly use properties of vector calculus.

$$f(\boldsymbol{\theta}) = (\boldsymbol{x} - \boldsymbol{\theta})^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{\theta}) \qquad \frac{df}{d\boldsymbol{\theta}} = 2(\boldsymbol{\theta} - \boldsymbol{x})$$

(using **vector** chain rule, **vector** power rule)

Important Properties

- **Linearity:**

$$\frac{d}{d\boldsymbol{x}}[a \cdot f(\boldsymbol{x}) + b \cdot g(\boldsymbol{x})] = a \cdot \frac{df}{d\boldsymbol{x}} + b \cdot \frac{dg}{d\boldsymbol{x}}$$

- **Product Rule:**

$$\frac{d}{d\boldsymbol{x}} f(\boldsymbol{x}) \cdot g(\boldsymbol{x}) = f(\boldsymbol{x}) \cdot \frac{dg}{d\boldsymbol{x}} + g(\boldsymbol{x}) \cdot \frac{df}{d\boldsymbol{x}}$$

$$\frac{d}{d\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) = \left(\frac{d\boldsymbol{g}}{d\boldsymbol{x}^{\mathrm{T}}}\right)^{\mathrm{T}} \boldsymbol{f}(\boldsymbol{x}) + \left(\frac{d\boldsymbol{f}}{d\boldsymbol{x}^{\mathrm{T}}}\right)^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x})$$

Jacobian

- **Chain Rule:**

$$\frac{df}{d\boldsymbol{x}} = \frac{df}{dg} \cdot \frac{dg}{d\boldsymbol{x}} = f'(g(\boldsymbol{x})) \cdot \frac{dg}{d\boldsymbol{x}}$$ $f: \mathbb{R} \to \mathbb{R},\ g: \mathbb{R}^p \to \mathbb{R}$

$$\frac{df}{d\boldsymbol{x}} = \left(\frac{d\boldsymbol{g}}{d\boldsymbol{x}^{\mathrm{T}}}\right)^{\mathrm{T}} \frac{df}{d\boldsymbol{g}}$$ $f: \mathbb{R}^q \to \mathbb{R},\ \boldsymbol{g}: \mathbb{R}^p \to \mathbb{R}^q$

(p x 1) (p x q) (q x 1) Hint: the sizes should match up.

Examples

$$L = \boldsymbol{x}^{\mathrm{T}}\mathbf{A}\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{x})$$

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}, \qquad \boldsymbol{g}(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x}$$

$$\frac{dL}{d\boldsymbol{x}} = \left(\frac{d\boldsymbol{f}}{d\boldsymbol{x}^{\mathrm{T}}}\right)^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) + \left(\frac{d\boldsymbol{g}}{d\boldsymbol{x}^{\mathrm{T}}}\right)^{\mathrm{T}} \boldsymbol{f}(\boldsymbol{x})$$

$$\frac{dL}{d\boldsymbol{x}} = \mathbf{A}\boldsymbol{x} + \mathbf{A}^{\mathrm{T}}\boldsymbol{x} = \left(\boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}}\right)\boldsymbol{x}$$

Examples

$$L = \boldsymbol{x}^{\mathrm{T}}\mathbf{A}\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{x})$$

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}, \qquad \boldsymbol{g}(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x}$$

$$\frac{dL}{d\boldsymbol{x}} = \left(\frac{d\boldsymbol{f}}{d\boldsymbol{x}^{\mathrm{T}}}\right)^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) + \left(\frac{d\boldsymbol{g}}{d\boldsymbol{x}^{\mathrm{T}}}\right)^{\mathrm{T}} \boldsymbol{f}(\boldsymbol{x})$$

$$\frac{dL}{d\boldsymbol{x}} = \mathbf{A}\boldsymbol{x} + \mathbf{A}^{\mathrm{T}}\boldsymbol{x} = \left(\boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}}\right)\boldsymbol{x}$$

$$L = f(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}) = f\left(g(\boldsymbol{x})\right)$$

$$g(\boldsymbol{x}) = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$$

$$\frac{dL}{d\boldsymbol{x}} = \frac{df}{dg} \cdot \frac{dg}{d\boldsymbol{x}} = f'(g) \cdot \boldsymbol{a} = f'\left(\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}\ \right) \cdot \boldsymbol{a}$$

Other Important Properties

- Refer to **The Matrix Cookbook**

$$\boxed{\frac{\partial \mathbf{x}^T\mathbf{B}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{B}+\mathbf{B}^T)\mathbf{x}} \tag{81}$$

$$\frac{\partial \mathbf{b}^T\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{c}}{\partial \mathbf{X}} = \mathbf{D}^T\mathbf{X}\mathbf{b}\mathbf{c}^T + \mathbf{D}\mathbf{X}\mathbf{c}\mathbf{b}^T \tag{82}$$

$$\frac{\partial}{\partial \mathbf{X}}(\mathbf{X}\mathbf{b}+\mathbf{c})^T\mathbf{D}(\mathbf{X}\mathbf{b}+\mathbf{c}) = (\mathbf{D}+\mathbf{D}^T)(\mathbf{X}\mathbf{b}+\mathbf{c})\mathbf{b}^T \tag{83}$$

Assume $\mathbf{W}$ is symmetric, then

$$\frac{\partial}{\partial \mathbf{s}}(\mathbf{x}-\mathbf{A}\mathbf{s})^T\mathbf{W}(\mathbf{x}-\mathbf{A}\mathbf{s}) = -2\mathbf{A}^T\mathbf{W}(\mathbf{x}-\mathbf{A}\mathbf{s}) \tag{84}$$

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}-\mathbf{s})^T\mathbf{W}(\mathbf{x}-\mathbf{s}) = 2\mathbf{W}(\mathbf{x}-\mathbf{s}) \tag{85}$$

- Refer to Wikipedia:
  - https://en.wikipedia.org/wiki/Gradient
  - https://en.wikipedia.org/wiki/Matrix_calculus

Next week: convexity, gradient descent, etc…