Machine Learning

4771

Instructor: Tony Jebara

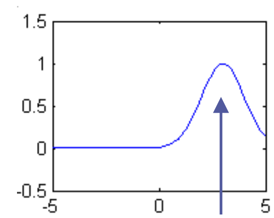
Topic 11

- Maximum Likelihood as Bayesian Inference
- Maximum A Posteriori
- Bayesian Gaussian Estimation

Why Maximum Likelihood?

- So far, assumed max (log) likelihood (IID or otherwise)
- Philosophical: Why?

$$\max_\theta L(\theta) = \max_\theta p(x_1,\ldots,x_N \mid \theta)$$
$$= \max_\theta \prod_{i=1}^{N} p(x_i \mid \theta)$$



- Also, why ignore p(θ)?

- Hint: Recall Bayes rule:

likelihood

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

prior

posterior

evidence

- Everyone agrees on probability theory: inference and use of probability models when we have computed p(x)
- But how get to p(x) from data? Debate…
- Two schools of thought: Bayesians and Frequentists

Bayesians & Frequentists

•Frequentists (Neymann/Pearson/Wald). An orthodox view that sampling is infinite and decision rules can be sharp.

•Bayesians (Bayes/Laplace/de Finetti). Unknown quantities are treated probabilistically and the state of the world can always be updated.

de Finetti: p(event) = price I would pay for a contract that pays 1$ when event happens

•Likelihoodists (Fisher). Single sample inference based on maximizing the likelihood function and relying on the Birnbaum's Theorem. Bayesians – But they don't know it.

Bayesians & Frequentists

- Frequentists:
  - Data are a repeatable random sample- there is a frequency
  - Underlying parameters remain constant during this repeatable process
  - Parameters are fixed

- Bayesians:
  - Data are observed from the realized sample.
  - Parameters are unknown and described probabilistically
  - Data are fixed

Bayesians & Frequentists

- Frequentists: classical / objective view / no priors
  every statistician should compute same p(x) so no priors
  can't have a p(event) if it never happened
  avoid $p(\theta)$, there is 1 true model, not distribution of them
  permitted: $p_\theta(x,y)$ forbidden: $p(x,y|\theta)$
  Frequentist inference: estimate one best model θ
    - use the ML estimator (unbiased & minimum variance)
    - do not depend on Bayes rule for learning
- Bayesians: subjective view / priors are ok
  put a distribution or pdf on all variables in the problem
  even models & deterministic quantities (i.e. speed of light)
  use a prior $p(\theta)$, on the model θ before seeing any data
  Bayesian inference: use Bayes rule for learning, integrate over all model (θ) unknown variables

Bayesian Inference

- Bayes rule gives rise to maximum likelihood
- Assume we have a prior over models p(θ)

likelihood

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

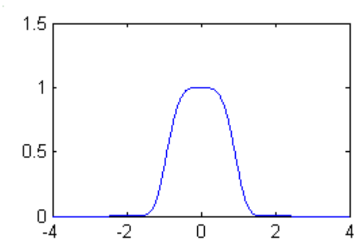
prior

posterior

evidence

- How to pick p(θ)?
  Pick simpler θ is better
  Pick form for mathematical convenience

$p\left(\theta\right)$



- We have data (can assume IID): $\mathfrak{X} = \left\{x_1, x_2, \ldots, x_N\right\}$
- Want to get a model to compute: $p\left(x\right)$
- Want p(x) given our data… How to proceed?

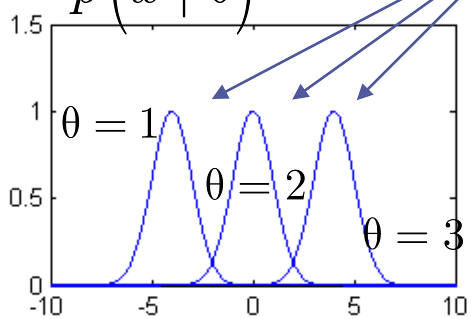
Bayesian Inference

- Want p(x) given our data… $p\left(x \mid \mathcal{X}\right) = p\left(x \mid x_1, x_2, \ldots, x_n\right)$

$$
\begin{aligned}
p\left(x \mid \mathcal{X}\right) &= \int_\theta p\left(x, \theta \mid \mathcal{X}\right) d\theta \\
&= \int_\theta p\left(x \mid \theta, \mathcal{X}\right) p\left(\theta \mid \mathcal{X}\right) d\theta \\
&= \int_\theta p\left(x \mid \theta, \mathcal{X}\right) \frac{p\left(\mathcal{X} \mid \theta\right) p\left(\theta\right)}{p\left(\mathcal{X}\right)} d\theta \\
&= \int_\theta p\left(x \mid \theta\right) \frac{\prod_{i=1}^{N} p\left(x_i \mid \theta\right) p\left(\theta\right)}{p\left(\mathcal{X}\right)} d\theta
\end{aligned}
$$

Prior

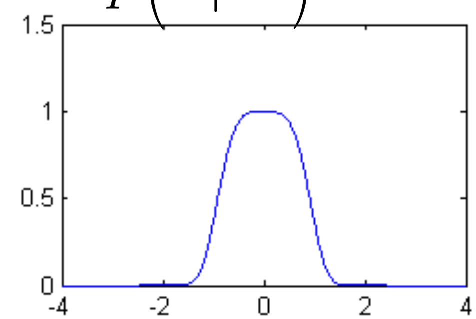
$p\left(x \mid \theta\right)$



Many models

Weight on each model

$p\left(\theta \mid \mathcal{X}\right)$



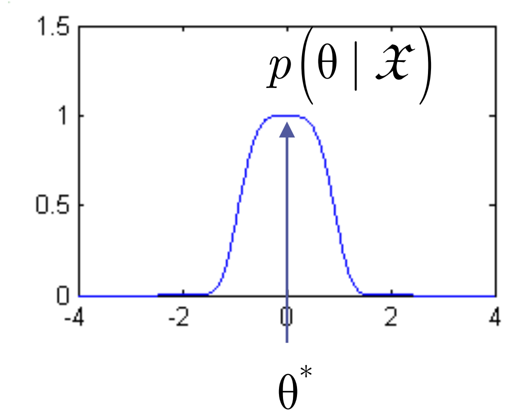
Bayesian Inference to MAP & ML

- The full Bayesian Inference integral can be mathematically tricky. Maximum likelihood is an approximation of it...

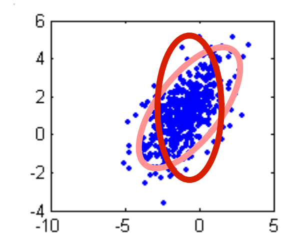
$$p(x \mid \mathcal{X}) = \int_\theta p(x \mid \theta) \frac{\prod_{i=1}^N p(x_i \mid \theta) p(\theta)}{p(\mathcal{X})} d\theta$$

$$\approx \int_\theta p(x \mid \theta) \delta(\theta - \theta^*) d\theta$$

$$where\ \theta^* = \begin{cases} \arg\max_\theta \frac{\prod_{i=1}^N p(x_i \mid \theta) p(\theta)}{p(\mathcal{X})} & MAP \\ \arg\max_\theta \frac{\prod_{i=1}^N p(x_i \mid \theta) uniform(\theta)}{p(\mathcal{X})} & ML \end{cases}$$



- Maximum A Posteriori (MAP) is like Maximum Likelihood (ML) with a prior p(θ) which lets us prefer some models over others



$$l_{MAP}(\theta) = l_{ML}(\theta) + \log p(\theta) = \sum_{i=1}^N \log p(x_i \mid \theta) + \log p(\theta)$$

Bayesian Inference Example

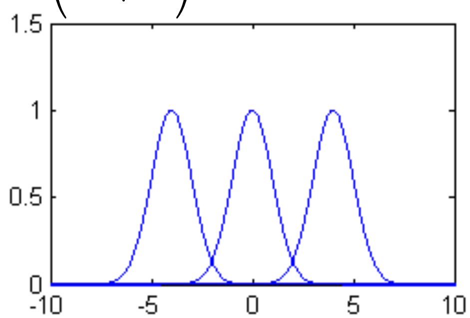
- For Gaussians, we CAN compute the integral (but hard!)

$$p\left(x \mid \mathcal{X}\right) = \int_\theta p\left(x \mid \theta\right) \frac{\prod_{i=1}^N p\left(x_i \mid \theta\right) p\left(\theta\right)}{p\left(\mathcal{X}\right)} d\theta$$

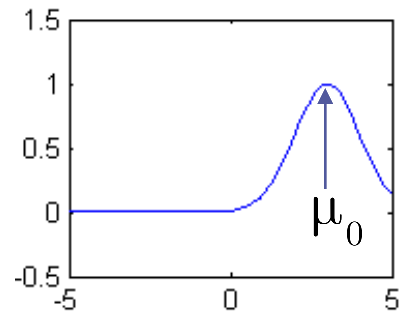
$$\propto \int_\theta p\left(x \mid \theta\right) \prod_{i=1}^N p\left(x_i \mid \theta\right) p\left(\theta\right) d\theta$$

- Example:... assume 1d Gaussian & Gaussian prior on mean

$$p\left(x \mid \theta\right) = Gaussian$$



$$p\left(\theta\right) = Gaussian$$



$$p\left(x \mid \mathcal{X}\right) \propto \int_\mu \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(x-\mu\right)^2} \right) \prod_{i=1}^N \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(x_i-\mu\right)^2} \right) \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\mu-\mu_0\right)^2} \right) d\mu$$

Bayesian Inference Example

- Solve integral over all Gaussian means with variance=1

$$p\left(x \mid \mathcal{X}\right) \propto \int_{\mu=-\infty}^{\mu=\infty} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(x-\mu\right)^2}\right) \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(x_i-\mu\right)^2}\right) \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\mu_0-\mu\right)^2}\right) d\mu$$

$$\propto \int_{\mu=-\infty}^{\mu=\infty} \exp\left(-\frac{1}{2}\left(x-\mu\right)^2 - \sum_i \frac{1}{2}\left(x_i-\mu\right)^2 - \frac{1}{2}\left(\mu_0-\mu\right)^2\right) d\mu$$

$$\propto \int_{\mu=-\infty}^{\mu=\infty} \exp\left(-\frac{1}{2}\left[\left(N+2\right)\mu^2 - 2\mu\left(x+\mu_0+\sum_i x_i\right) + x^2\right]\right) d\mu$$

$$\propto \int_{\mu=-\infty}^{\mu=\infty} \exp\left(-\frac{1}{2}\left[\left(N+2\right)\mu^2 - 2\mu\left(x+\mu_0+\sum_i x_i\right) + x^2\right] + \left[\quad\right]^2 - \left[\quad\right]^2\right) d\mu$$

$$\propto \exp\left(-\frac{1}{2}\left[\frac{-\left(x+\mu_0+\sum_i x_i\right)^2}{N+2} + x^2\right]\right)$$

$$\tilde{\mu} = \frac{\mu_0 + \sum_i x_i}{N+1}$$

$$= N\left(x \mid \tilde{\mu}, \tilde{\sigma}^2\right) \qquad \tilde{\sigma}^2 = \frac{N+2}{N+1}$$

- Can integrate over μ and Σ for multivariate Gaussian (Jordan ch. 4 and Minka Tutorial)

$$p\left(x \mid \mathcal{X}\right) = \frac{\Gamma\left(\left(N+1\right)/2\right)}{\Gamma\left(\left(N+1-d\right)/2\right)} \left|\frac{1}{\left(N+1\right)\pi}\bar{\Sigma}^{-1}\right|^{1/2} \left(\frac{1}{N+1}\left(x-\bar{\mu}\right)^T \bar{\Sigma}^{-1}\left(x-\bar{\mu}\right)+1\right)^{-\left(N+1\right)/2}$$