Lecture 3 Bayesian Parameter Estimation

problems w/ MLE

- Model a coin flip w/ Bernoulli r.v. $\{0=T, 1=H\}$
- MLE: $\hat{\pi} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- Suppose we see $D = \{1,1,1,0,0,0,1\} \Rightarrow \hat{\pi} = \frac{4}{7}$ ✓
- Suppose we see $D = \{1,1,1\} \Rightarrow \hat{\pi} = \frac{3}{3} = 1$

  $\pi = p(1) = 1$

  $1-\pi = p(0) = 0$ ← probability 0 of happening!

  (This is unreasonable → we can never see tails!)

- This is an example of overfitting.
- We know that coins are usually fair ($\pi = \frac{1}{2}$).

  How to incorporate this knowledge into our estimator?

CS5487 Lecture Notes (2020)
Dr. Antoni B. Chan
Dept of Computer Science
City University of Hong Kong

Bayesian Parameter Estimation

- treat parameters θ as a r.v.
- Framework
  - training set: $D = \{x_1, \ldots, x_N\}$
  - likelihood of data given parameter $p(x_i|\theta)$
  - prior distribution on parameters $p(\theta)$ (encode our beliefs/knowledge about θ)
- posterior distribution of θ given the data D

$$p(\theta|D) = \frac{p(D|\theta)\,p(\theta)}{p(D)}, \quad p(D) = \int p(D|\theta)p(\theta)\,d\theta$$

(computing a distribution rather than a point estimate, like MLE)

- predictive distribution - likelihood of a new sample x_*, given the data D.

$$p(x_*|D) = \int p(x_*|\theta)\,p(\theta|D)\,d\theta$$

(averaging over all θ weighted by the posterior)

"allow different explanations of the data"

Example: Gaussian (known variance)

prior: $p(\mu) = N(\mu \mid \mu_0, \sigma_0^2)$ ← known (prior beliefs)

observation likelihood: $p(x \mid \mu) = N(x \mid \mu, \sigma^2)$ ← known

Dataset: $D = \{x_1, \dots, x_N\}$

posterior: $$p(\mu \mid D) = \frac{p(D \mid \mu)\, p(\mu)}{p(D)}$$

$$= \frac{\prod_{i=1}^{N} p(x_i \mid \mu)\; p(\mu)}{\int \prod_{i=1}^{N} p(x_i \mid \mu)\, p(\mu)\, d\mu}$$ ← doesn't depend on μ

- Look at the numerator first and see the form in terms of μ of the distribution, then normalize later.

$$\log p(\mu \mid D) \propto \sum_{i=1}^{N} \log p(x_i \mid \mu) + \log p(\mu)$$

$$= \left[\sum_{i=1}^{N} -\frac{1}{2\sigma^2}(x_i - \mu)^2 - \frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2\right] - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2 - \frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma_0^2$$

$$\propto \sum_{i=1}^{N} -\frac{1}{2\sigma^2}(x_i^2 - 2x_i\mu + \mu^2) - \frac{1}{2\sigma_0^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2)$$

$$\propto -\frac{1}{2}\left[\frac{-2}{\sigma^2}\underbrace{\sum_i x_i}_{N\hat{\mu}_{ML}}\mu + \frac{1}{\sigma^2}N\mu^2 + \frac{1}{\sigma_0^2}\mu^2 - \frac{2}{\sigma_0^2}\mu\mu_0\right]$$

$$= -\frac{1}{2}\left[\underbrace{\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)}_{a}\mu^2 - 2\underbrace{\left(\frac{N}{\sigma^2}\hat{\mu}_{ML} + \frac{1}{\sigma_0^2}\mu_0\right)}_{b}\mu\right]$$

$$a\mu^2 - 2b\mu$$

$$= a(\mu - d)^2 + e$$

completing the square: $d = b/a$

$$e = -\frac{b^2}{a} = -ad^2$$

$$= -\frac{1}{2}\frac{1}{\hat{\sigma}_n^2}\left(\mu - \hat{\mu}_n\right)^2 + \text{const}$$

where

$$\hat{\sigma}_n^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}$$

$$\hat{\mu}_n = \left[\frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}\right]\cdot\left(\frac{N}{\sigma^2}\hat{\mu}_{ML} + \frac{1}{\sigma_0^2}\mu_0\right)$$

$$= \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}\left[\frac{N}{\sigma^2}\hat{\mu}_{ML} + \frac{1}{\sigma_0^2}\mu_0\right]$$

$$\hat{\mu}_n = \frac{N\sigma_0^2\,\hat{\mu}_{ML} + \sigma^2\mu_0}{N\sigma_0^2 + \sigma^2}$$

Since the $\log p(\mu|D)$ is quadratic in μ

$\Rightarrow$ Gaussian

$$p(\mu|D) = N(\mu \mid \hat{\mu}_n, \hat{\sigma}_n^2)$$

let $\alpha = \frac{N\sigma_0^2}{\sigma^2 + N\sigma_0^2}$

$$\hat{\mu}_n = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\hat{\mu}_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 = \alpha\hat{\mu}_{ML} + (1-\alpha)\mu_0$$

(α on the first term, $1-\alpha$ on the second)

$$\hat{\sigma}_n^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}}$$

What's the interpretation?

Data size:

$N=0 \Rightarrow \alpha=0 \Rightarrow \hat{\mu}_n = \mu_0$ (no data, use prior)

$N\to\infty \Rightarrow \alpha=1 \Rightarrow \hat{\mu}_n = \hat{\mu}_{ML}$ (lots of data, use MLE)

* smoothing between MLE & prior.

$N=0 \Rightarrow \hat{\sigma}_n^2 = \sigma_0^2$ (use prior uncertainty)

$N\to\infty \Rightarrow \hat{\sigma}_n^2 = 0$ (converges to a single value, the MLE)

$\sigma_0^2 \ll \sigma^2 \Rightarrow \alpha=0 \Rightarrow \hat{\mu}_n = \mu_0$ (observations too noisy $\Rightarrow$ use prior)

$\sigma_0^2 \gg \sigma^2 \Rightarrow \alpha=1 \Rightarrow \hat{\mu}_n = \hat{\mu}_{ML}$ (weak belief, use MLE)

$\sigma^2 = \sigma_0^2 \Rightarrow \alpha = \frac{N}{N+1} \Rightarrow$

$$\hat{\mu}_N = \frac{N}{N+1}\hat{\mu}_{ML} + \frac{1}{N+1}\mu_0$$

$$= \frac{1}{N+1}\left(N\hat{\mu}_{ML} + \mu_0\right)$$

$$= \frac{1}{N+1}\left(\sum_{i=1}^{N} x_i + \mu_0\right)$$

- adding 1 "virtual" sample at μ_0, then compute the mean.
- for small N, the virtual sample has an effect
- for large N, it has a little effect.

This is a form of regularization.

Predictive distribution

$$p(\mu|D) = N(\mu|\hat{\mu}_n, \hat{\sigma}_n^2)$$
$$p(x|\mu) = N(x|\mu, \sigma^2)$$

$$p(x|D) = \int \underbrace{N(x|\mu, \sigma^2)}_{N(\mu|x,\sigma^2)} N(\mu|\hat{\mu}_n, \hat{\sigma}_n^2)\, d\mu$$

→ multiply 2 Gaussians (PS1-7)

$$= \int N(x|\hat{\mu}_N, \sigma^2 + \hat{\sigma}_n^2)\, N(\mu|\ldots\ldots)\, d\mu$$

$$\boxed{p(x|D) = N(x|\hat{\mu}_n, \sigma^2 + \hat{\sigma}_n^2)}$$

- $\hat{\mu}_n$: same as posterior
- σ^2: observation noise
- $\hat{\sigma}_n^2$: uncertainty in μ

Maximum a Posteriori (MAP)

With full Bayesian estimation we need
$p(D) = \int p(D|\theta)p(\theta)d\theta$ ... usually hard to do.

one solution: pick θ w/ highest posterior probability

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\ p(\theta|D)$$
$$= \underset{\theta}{\operatorname{argmax}}\ \frac{p(D|\theta)p(\theta)}{p(D)}$$
$$= \underset{\theta}{\operatorname{argmax}}\ p(D|\theta)p(\theta)$$

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\ \log p(D|\theta) + \log p(\theta)$$

- $\log p(D|\theta)$: data log-likelihood (like MLE)
- $\log p(\theta)$: prior (regularize the θ)

Example: Gaussian

$$\hat{\mu}_{MAP} = \underset{\mu}{\operatorname{argmax}}\ p(\mu|D) = \hat{\mu}_n$$

predictive distribution: $p(x|\hat{\mu}_{MAP}) = N(x|\hat{\mu}_n, \sigma^2)$

Bayesian Regression

similar to before

α = variance of the prior (known)

function: $f(x,\theta) = \phi(x)^T\theta$

noisy obs: $y = f(x,\theta) + \epsilon$, $\epsilon \sim N(0, \sigma^2)$

* prior on θ: $p(\theta) = N(\theta \mid 0, \alpha I)$ — I = identity matrix

$$\alpha I = \begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix}$$

data likelihood: $p(y \mid x, \theta) = N(y \mid f(x), \sigma^2)$

MAP estimate θ

$$\hat{\theta} = \arg\max_\theta \log p(D \mid \theta) + \log p(\theta)$$

$$= \arg\max_\theta \sum_{i=1}^{N} \log p(y_i \mid x_i, \theta) + \log p(\theta)$$

same as tutorial

$$= \arg\max_\theta -\frac{1}{2\sigma^2}\|y - \Phi^T\theta\|^2 - \frac{1}{2}\frac{1}{\alpha}\|\theta\|^2 + \text{const.}$$

$$= \arg\min_\theta \left(\frac{1}{\sigma^2}\|y - \Phi^T\theta\|^2 + \frac{1}{\alpha}\|\theta\|^2 \right) \cdot \sigma^2$$

$$\hat{\theta}_{MAP} = \arg\min_\theta \|y - \Phi^T\theta\|^2 + \lambda\|\theta\|^2 \qquad \frac{\sigma^2}{\alpha} = \lambda$$

λ: hyperparameter — controls the length of θ ⇒ controls the complexity of f.

$f(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots \theta_p x^p$

"ridge regression"
"regularized L.S."
"Tikhonov regularization"
"shrinkage"
"weight decay"

$$\hat{\theta} = (\Phi\Phi^T + \lambda I)^{-1}\Phi y.$$

L.S: $\hat{\theta} = (\Phi\Phi^T)^{-1}\Phi y$

λI regularizes the covariance matrix $\Phi\Phi^T$.
prevents ill-conditioned matrix & problems of inverting it.

"ridge"

Tutorial 3

PS 3.10 Bayesian Regression w/ Gaussian Prior.

parameter vector $\theta \in \mathbb{R}^p$

$y = \Phi^T \theta + \epsilon$ ← Noise $\epsilon \sim N(0, \Sigma)$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \qquad \Phi = \begin{bmatrix} 1 & & 1 \\ \phi(x_1) & \dots & \phi(x_N) \\ 1 & & 1 \end{bmatrix}$$

Prior: $\theta \sim N(0, \Gamma)$

observation likelihood: $p(y|X,\theta) = N(y | \Phi^T\theta, \Sigma)$

a) Posterior of $\theta | X, y$

$$p(\theta|X,y) = \frac{p(y|X,\theta)\,p(\theta)}{p(y|X)}$$ (Bayes Rule)

Look at the log of numerator to see the form of θ:

$$\log p(y|X,\theta) + \log p(\theta)$$
$$= -\frac{1}{2}\|y - \Phi^T\theta\|^2_\Sigma - \frac{1}{2}\log|\Sigma| - \frac{N}{2}\log 2\pi$$
$$- \frac{1}{2}\|\theta\|^2_\Gamma - \frac{1}{2}\log|\Gamma| - \frac{p}{2}\log 2\pi$$

$$\propto -\frac{1}{2}\left[y^T\Sigma^{-1}y - 2y^T\Sigma^{-1}\Phi^T\theta + \theta^T\Phi\Sigma^{-1}\Phi^T\theta + \theta^T\Gamma^{-1}\theta \right]$$

$$\propto -\frac{1}{2}\left[\theta^T\underbrace{(\Phi\Sigma^{-1}\Phi^T + \Gamma^{-1})}_{A}\theta - 2\underbrace{y^T\Sigma^{-1}\Phi^T}_{b^T}\theta \right]$$

$$\theta^TA\theta - 2b^T\theta \Rightarrow (\theta - \hat{\mu})^T\hat{\Sigma}^{-1}(\theta - \hat{\mu})$$

P1.10 complete the square

$$f(x) = x^TAx - 2b^Tx + c$$
$$= (x-d)^TA(x-d) + e$$

where $d = A^{-1}b$

$e = c - b^TA^{-1}b$

$$\hat{\mu} = \underbrace{(\Phi\Sigma^{-1}\Phi^T + \Gamma^{-1})^{-1}}_{A^{-1}}\underbrace{\Phi\Sigma^{-1}y}_{b}$$

$$\hat{\Sigma} = \underbrace{(\Phi\Sigma^{-1}\Phi^T + \Gamma^{-1})^{-1}}_{A^{-1}}$$

$$\Rightarrow \log p(y|x,\theta)p(\theta) \propto -\frac{1}{2}(\theta - \hat{\mu})^T\hat{\Sigma}^{-1}(\theta - \hat{\mu})$$

since the log numerator is quadratic, then it must be Gaussian (constants will combine to form the Gaussian normalization $\frac{1}{(2\pi)^{N/2}|\hat{\Sigma}|^{1/2}}$)

$$\Rightarrow p(\theta|X,y) = N(\theta | \hat{\mu}, \hat{\Sigma})$$

b) MAP estimate

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\; p(\theta|X,y) =$$
$$= \hat{\mu} = (\Phi\Sigma^{-1}\Phi^T + \Gamma^{-1})\Phi\Sigma^{-1}y$$

$$\hat{\theta}_{MAP} = \left(\Phi \Sigma^{-1} \Phi^T + \Gamma^{-1}\right)^{-1} \Phi \Sigma^{-1} y$$

regularizer - helps make the inverse better conditioned.

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_N^2 \end{bmatrix} \Rightarrow \Sigma^{-1} = \begin{bmatrix} 1/\sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & 1/\sigma_N^2 \end{bmatrix}$$

$$\Phi \Sigma^{-1} = \begin{bmatrix} \phi_1 & \cdots & \phi_N \end{bmatrix} \begin{bmatrix} 1/\sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & 1/\sigma_N^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_1^2}\phi_1 & \cdots & \frac{1}{\sigma_N^2}\phi_N \end{bmatrix}$$

large observation noise $\sigma_i^2 \Rightarrow$ lower weight on this point (y_i, x_i)

$\Rightarrow$ Similar to weighted least squares.

$\Rightarrow$ similar to regularized least squares

Bayesian regression combines them.

c) assume $\Gamma = \alpha I$, $\Sigma = \sigma^2 I$

$$\hat{\theta}_{MAP} = \left(\Phi \Sigma^{-1} \Phi^T + \Gamma^{-1}\right)^{-1} \Phi \Sigma^{-1} y$$

$$= \left(\Phi \frac{1}{\sigma^2} I \Phi^T + \frac{1}{\alpha} I\right)^{-1} \Phi \frac{1}{\sigma^2} I y$$

(σ^2)

$$= \left(\Phi \Phi^T + \frac{\sigma^2}{\alpha} I\right)^{-1} \Phi y$$

(λ)

$$\hat{\theta}_{MAP} = \left(\Phi \Phi^T + \lambda I\right)^{-1} \Phi y$$

same as regularized least squares!

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|y - \Phi^T \theta\|^2 + \lambda \|\theta\|^2$$

$$= \underset{\theta}{\operatorname{argmin}} \; y^T y - 2 y^T \Phi^T \theta + \theta^T \Phi \Phi^T \theta + \lambda \theta^T \theta$$

$$\frac{\partial}{\partial \theta} = -2 \Phi y + \underbrace{2 \Phi \Phi^T \theta + 2\lambda \theta}_{2(\Phi \Phi^T + \lambda I)\theta} = 0$$

$$=$$

$$\hat{\theta} = \left(\Phi \Phi^T + \lambda I\right)^{-1} \Phi y$$

same!

e) predictive distribution

Given an x_*, $f(x_*)$

$$f_* = f(x_*) = \theta^T \phi(x_*)$$

Note $\theta \sim p(\theta|X,y) = N(\theta|\hat{\mu}, \hat{\Sigma})$

$$p(f_*|X,y,x_*)$$

xformations of r.v. (p1.1)

$y = Ax+b$, x is r.v.

$E[y] = A E[x] + b$

$cov(y) = A\, cov(x)\, A^T$

f_* is a r.v. that is a linear xform of $\theta|X,y$

$\Rightarrow E[f_*] = \phi(x_*)^T \hat{\mu}$

$\Rightarrow cov(f_*) = \phi(x_*)^T \hat{\Sigma} \phi(x_*)$

Since θ is Gaussian, then

$$f_*|X,y,x_* \sim N(f_*|\phi(x_*)^T\hat{\mu}, \phi(x_*)^T\hat{\Sigma}\phi(x_*))$$