motivation problem of MLE

Coin flipping Bernoulli r.v. $-\{0=T, 1=H\}$

MLE: $\hat{\pi} = \frac{1}{N}\sum_i x_i$

Suppose we see: $D = \{1,1,1,0,0,0,0\} \Rightarrow \hat{\pi} = \frac{3}{7}$

what if $D' = \{1,1,1\}$ only $\hat{\pi} = 1$? (we never see tails)

This is an example of overfitting. (not enough samples to get a good estimate of the parameter)

What we can do?

- use our knowledge: we know $\pi \approx \frac{1}{2}$ for most coins and we incorperate this knowledge to our estimate of π.

Bayesian Param Estimation

- treat θ as a r.v.
- Framework
  - training set $D = \{x_1, \ldots\ldots, x_N\}$
  - prob density given parameter θ: $p(x_i|\theta)$
  - prior distribution on parameter θ, $p(\theta)$ (added)
    (encode prior beliefs about θ, eg: $\pi \approx \frac{1}{2}$)
- posterior dist. of θ given data D.

$$\underbrace{p(\theta|D)}_{\text{updated}} = \frac{p(D|\theta)\cdot p(\theta)}{\int p(D|\theta)\, p(\theta)\, d\theta} \Rightarrow \text{density functions}$$

- predictive dist. – likelihood of new x_* given data D.

$$p(x_*|D) = \underbrace{\int p(x_*|\theta)\, p(\theta|D)\, d\theta}_{\text{average over all } \theta,\ \text{weighted by posterior } p(\theta|D)}$$

how likely is the θ (weight)
already fixed in the formula.

"allow different explanations of data"

compared to ML, Bayes is influenced by prior

pure determined by data

$\Rightarrow$ (problem: how to get prior)

Example: Gaussian (known variance)

prior on μ: $p(\mu) = \mathcal{N}(\mu \mid \mu_0, \alpha_0^2)$

prior are given

Likelihood of x: $p(x \mid \mu) = \mathcal{N}(x \mid \mu, \alpha^2)$

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

Calculate posterior

$$p(\mu \mid D) = \frac{[\prod_i p(x_i \mid \mu)]\, p(\mu)}{\int [\prod_i p(x_i \mid \mu)]\, p(\mu)\, d\mu}$$

product of gaussian

doesn't depend on μ.

☆ Just look at numerator w.r.t. μ, then normalize later.

product of Gaussian

can swap $(x-\mu)^2 = (\mu - x)^2$

$$\mathcal{N}(x \mid a, A) \cdot \mathcal{N}(x \mid b, B) = \mathcal{N}(a \mid b, A+B)\, \mathcal{N}(x \mid c, C)$$

$$C = \frac{1}{\frac{1}{A} + \frac{1}{B}} \Rightarrow \frac{1}{C} = \frac{1}{A} + \frac{1}{B} \qquad c = C\left(\frac{a}{A} + \frac{b}{B}\right)$$

first 2 terms:

$$p(x_1 \mid \mu) \cdot p(x_2 \mid \mu) = \mathcal{N}(\mu \mid x_1, \alpha^2) \cdot \mathcal{N}(\mu \mid x_2, \alpha^2)$$

$$= \mathcal{N}(x_1 \mid x_2, 2\alpha^2)\, \mathcal{N}(\mu \mid \tilde{\mu}_2, \tilde{\alpha}_2^2)$$

$$\begin{cases} \frac{1}{\tilde{\alpha}_2^2} = \frac{1}{\alpha^2} + \frac{1}{\alpha^2} = \frac{2}{\alpha^2} \\ \tilde{\mu}_2 = \frac{\alpha^2}{2}\left(\frac{x_1}{\alpha^2} + \frac{x_2}{\alpha^2}\right) = \frac{1}{2}(x_1 + x_2) \end{cases}$$

$\Rightarrow p(x_1 \mid \mu)\, p(x_2 \mid \mu) \propto \mathcal{N}(\mu \mid \tilde{\mu}_2, \tilde{\alpha}_2^2)$ (throw away the constant)

first 3 terms

$$\mathcal{N}(\mu \mid \tilde{\mu}_2, \tilde{\alpha}_2^2)\, \mathcal{N}(x_3 \mid \mu, \alpha^2) \propto \mathcal{N}(\mu \mid \tilde{\mu}_3, \tilde{\alpha}_3^2)$$

precision increasing $\Leftarrow$ $\frac{1}{\tilde{\sigma}_3^2} = \frac{1}{\tilde{\sigma}_2^2} + \frac{1}{\sigma^2} = \frac{2}{\sigma^2} + \frac{1}{\sigma^2} = \frac{3}{\sigma^2}$

$$\tilde{\mu}_3 = \frac{\sigma^2}{3}\left(\frac{\frac{1}{2}(x_1+x_2)}{\frac{\sigma^2}{2}} + \frac{x_3}{\sigma^2} \right) = \frac{1}{3}(x_1+x_2+x_3)$$

$\Rightarrow$ reduction

first N terms

$$\prod_i p(x_i|\mu) \propto \mathcal{N}(\mu|\tilde{\mu}_n, \tilde{\sigma}_n)$$

$$\begin{cases} \tilde{\mu}_n = \frac{1}{N}\sum_i x_i = \boxed{\hat{\mu}_{ML}} \\ \sigma_\mu^2 = \frac{\sigma^2}{N} \end{cases}$$

turn out to be MLE res.

Add prior

mul constant

$$\mathcal{N}(\mu|\tilde{\mu}_n, \tilde{\sigma}_n^2)\,\mathcal{N}(\mu|\mu_0, \sigma_0^2) \propto \mathcal{N}(\mu|\hat{\mu}_n, \hat{\sigma}_n^2)$$

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

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

mul $\frac{\sigma_0^2\sigma^2}{\sigma_0^2\sigma^2}$

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

Finally, $P(\mu|D) = \mathcal{N}(\mu|\hat{\mu}_n, \hat{\sigma}_\mu^2)$

☆ two constants $\Rightarrow$ cancel

What does it mean?

interpret between MLE sol and prior μ_0

Data Size

mean $N=0 \Rightarrow \alpha=0 \Rightarrow \hat{\mu}_n=\mu_0$

$N \to \infty \Rightarrow \alpha=1 \Rightarrow \hat{\mu}_n = \mu_{ML}$.

variance $N=0 \Rightarrow \hat{\sigma}_n^2 = \sigma_0^2$

$N \to \infty \Rightarrow \hat{\sigma}_n^2 \to 0 \leftarrow$ Converge to single value.

$\sigma_0^2 \ll \sigma^2 \Rightarrow \alpha=0 \Rightarrow \hat{\mu}_n=\mu_0$

small prior

(strong belief compared to noise $\Rightarrow$ use our belief)

$\sigma_0^2 \gg \sigma^2 \Rightarrow \alpha=1 \Rightarrow \hat{\mu}_n=\hat{\mu}_{ML}$

(weak belief $\to$ use MLE.)

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

$$= \underbrace{\frac{1}{N+1}\left(\sum_i x_i+\mu_0\right)}_{}$$

add a virtual sample at μ_0, then compute the mean

- for large N, the v.s. doesn't matter
- for small N, move the posterio towards μ_0

[This is a form of regularization]

predictive distribution.

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

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

$$p(x|D) = \int p(x|\mu) \cdot p(\mu|D) d\mu$$

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

$$= \int \mathcal{N}(x|\hat{\mu}_N, \sigma^2 + \hat{\sigma}_n^2) \underbrace{\mathcal{N}(\mu|\ldots,\ldots) d\mu}_{\text{normalized to } 1}$$

integration over μ

$$p(x|D) = \mathcal{N}(x|\underbrace{\hat{\mu}_N}_{\text{same mean as posterior}}, \hat{\sigma}_n^2 + \sigma^2)$$

variance of paramet- w/D (uncertainty)

due to noise observation (uncertainty)

Maximize a Posteriori (MAP)

Avoid calculating the denominator of Bayes' Rule

$$\int p(D|\theta) p(\theta) d\theta \text{ ---- difficult.}$$

Solu. pick the θ with largest posterior prob.

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|D)$$

$$= \arg\max_{\theta} \frac{p(D|\theta) p(\theta)}{\int p(D|\theta) \cdot p(\theta) d\theta)} \Rightarrow \text{constant w.r.t. } \theta$$

$$= \arg\max p(D|\theta) p(\theta)$$

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \underbrace{\log p(D|\theta)}_{\text{data LL for MLE}} + \underbrace{\log p(\theta)}_{\text{regularization (belief)}}$$

Example Gaussian.

$$\hat{\mu}_{MAP} = \underset{\mu}{\operatorname{argmax}}\, p(\mu|D) = \underset{\mu}{\operatorname{argmax}}\, N(\mu|\hat{\mu}_n, \hat{\sigma}_n^2)$$

$$\hat{\mu}_{MAP} = \hat{\mu}_n$$

Approximate posterior as a delta function:

$$p(\mu|D) \approx \delta(\mu - \hat{\mu}_n)$$

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

Bayesian Regression

Same setup as before:

$$\left.\begin{array}{l} x \in \mathbb{R} \\ f(x) = \phi(x)^T\theta \\ y = f(x) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \end{array}\right] \Rightarrow p(y|x,\theta) = N(y|f(x), \sigma^2)$$

Introduce prior on θ:

$$p(\theta) = \mathcal{N}(\theta \mid 0, \alpha^2 I)$$

scaled identity covariance matrix

MAP estimate

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

$$= \arg\max_{\theta} \sum_i \log p(y_i \mid x_i, \theta) + \log p(\theta)$$

tutorial ⋮

$$= \arg\min_{\theta} \| y - \Phi^T \theta \|^2 + \lambda \|\theta\|^2$$

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

add some constant to eigenvalues

constant ← controls regularization

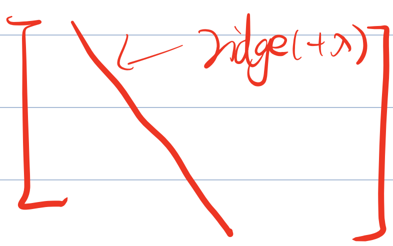
$\lambda = 0 \Rightarrow$ L.S.

(depend on α^2 and α)

ridge regression

- regularized least square
- shrinkage
- weight decay

regularize covariance matrix to prevent inverting an ill-conditioned matrix



(add λ to all the eigenvalues of $\Phi\Phi^T$)

(sidenote:

$$(A+\alpha I)x = ax$$

$$Ax^{*} = \lambda^{*} x^{*}$$

$$Ax + \alpha I x = \lambda x$$

$$Ax = (\lambda I - \alpha I)x$$