Lecture 3 Jan 28th

Estimation II

Process for finding unknown parameters that maximize likelihood of underlying distribution.

Common Assumptions:

- Each sample is independant
- Binomial θ is uniform(0,1).

Binomial II

$$P(C_i, \theta) = P(C_i|\theta)P(\theta) \propto P(\theta|C_i)P(C_i)$$

Thus

$$P(\theta|C_i) = \frac{P(C_i|\theta)P(\theta)}{P(C_i)}$$

So suppose $n = 1000$ with 750 heads.

$$P(C_i|\theta)P(\theta) = \prod P(C_i|\theta) = \theta^{750}(1-\theta)^{250}$$

And

$$P(C_i) = \int_0^1 P(C_i, \theta)\,d\theta = \int_0^1 \theta^{750}(1-\theta)^{1000-750}\,d\theta = Z$$

$$P(\theta|C_i) = \frac{1}{Z}\,\theta^{750}(1-\theta)^{1000-75}$$

$$\underset{\theta}{\text{Argmax}}\, P(\theta|C_i) = \frac{H}{N} \quad \begin{matrix} - \text{heads} \\ - n \end{matrix}$$

In General II

Use Baye's Rule:

$$\underbrace{p(\text{model}|\text{data})}_{\text{posterior}} = \frac{\overbrace{p(\text{data}|\text{model})}^{\text{likelihood}}\ \overbrace{p(\text{model})}^{\text{Prior}}}{\underbrace{p(\text{data})}_{\text{normalizing constant}}}$$

posterior evidence

Posterior Dist//
Knowledge of the model based on both the data and the prior.
Likelihood Dist //
Likelihood of sample data assuming model is correct.
Prior Dist//
Assumptions about the model prior to observing the data.
Evidence //
Used in model selection. Could normalize posterior PDF.

2 Ways of Choosing Estimators

Maximum A Posteriori
Let $\hat{\theta}$ be the optimal parameter; choose

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

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

Notice $P(D)$, the denominator in Bayes's Rule isn't included since it's often just a constant.

Maximum Likelihood
When no prior information is known about θ, we can ignore $P(\theta)$ term and choose,

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

Tip: Use the negative log function on MAP or ML to easily find the derivative of likelihood functions.

Gaussian Distribution (Normal)

These samples are distributed with $\vec{\mu} \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n \times n}$.

$$\sim N(\mu, \Sigma)$$

Thuse the likelihood func. is,

$$p(D \mid \mu, \Sigma) = \prod \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \cdot \exp\left(-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)\right)$$

Then the negative log-likelihood func is,

$$L(D \mid \mu, \Sigma) = \frac{n d}{2} \ln 2\pi + \frac{n}{2} \ln |\Sigma| + \frac{1}{2} \Sigma (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$$

By solving $\frac{dL}{d\mu} = 0$ and $\frac{dL}{d\Sigma}$, we get

$$\mu^* = \frac{1}{n} \Sigma x_i$$

$$\Sigma^* = \frac{1}{n} \Sigma (x_i - \mu^*)^T (x_i - \mu^*)^T$$

MAP Non-linear Regression

Now we use a more accurate model for non-linear regression; we account for noise in the data.

$$y = \vec{w}^T b(x) + n$$

where $\vec{w}$ is vector of weights, $\vec{b}(x)$ is vector with all basis functions and

$$n \sim N(0, \sigma^2)$$

Likelihood of Gaussian

Thus $E(y) = \vec{w}^T b(x)$

$Var(y) = \sigma^2$

$$p(\vec{y} \mid \vec{w}, \vec{x}) = \prod_i^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y_i - \vec{w}^T b(x)}{\sigma}\right)^2\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \cdot \exp\left(-\frac{1}{2\sigma^2}\sum (y_i - \vec{w}^T b(x))^2\right)$$

Let $\vec{w} \sim N(0, \alpha I)$, (assumption of normal)

$$p(\vec{w} \mid \vec{x}) = p(\vec{w}) = \prod_k^d \frac{1}{\sqrt{(2\pi)^d \alpha}} e\left(\frac{-w^2 k}{2\alpha}\right)$$

$$= \frac{1}{(2\pi\alpha)^{\frac{d}{2}}} e\left(-\frac{1}{2\alpha}\vec{w}^T\vec{w}\right)$$

Thus to optimize for $\vec{w}^*$.

$$p(\vec{w} \mid \vec{y}, \vec{x}) = \frac{p(\vec{y} \mid \vec{w}, \vec{x})\, p(\vec{w} \mid \vec{x})}{p(\vec{y} \mid \vec{x})}$$ by Bayes' Rule

Finally

$$L(\vec{w}) = \frac{n}{2}\ln 2\pi\sigma^2 + \frac{1}{2\sigma^2}\sum (y_i - \vec{w}^T b(x))^2 + \frac{d}{2}\ln(2\pi\alpha) + \frac{1}{2\alpha}\|\vec{w}\|^2 + \ln p(\vec{y} \mid \vec{x})$$

$$\propto \frac{1}{2\sigma^2}\sum (y_i - \vec{w}^T b(x)) + \frac{1}{2\alpha}\|\vec{w}\|^2$$

Since we don't care about the rest when we optimize / derive.

Note: above equation is exactly the same as the least squares method with regularization.