Bayesian Estimation.

Major diff from MLE : Parameter is assumed to be a RV with a dist.

Dist of Parameter RV : prior Knowledge of the parameter
[weight bias in the biased coin]

In MLE, final estimates are decided by data alone.

Let θ be the Parameter & D be the data.

$$D = \{x_1, x_2 \ldots x_n\}.$$

set of iid , Each $x_i \sim f(x_i/\theta)$.

let $f(\theta)$ be the prior of parameter & $f(\theta|D)$ be the posterior.

From Bayes theorem,

$$f(\theta|D) = \frac{f(D|\theta)\, f(\theta)}{\int f(D|\theta) f(\theta)\, d\theta}.$$

$$f(D|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$ – data likelihood as before.

Denominator is not a fn of θ & thus can be ignored.

Question: How to use $f(\theta|D)$ for the classifier?

We need class-cond for classifiers.

One choice: $f(x|D,y) = \int f(x,\theta|D)\,d\theta$

$= \int f(x|\theta)\, f(\theta|D)\,d\theta$

may get this depending upon the forms.

The other popular alternative: get point estimates of θ from $f(\theta|D)$

One choice: mode of $f(\theta|D)$ - called the MAP estimate.

One can use any measure of central tendency.

Eg:

Question 2: How to Choose the prior?

$$f(\theta|D) \propto f(D|\theta) f(\theta).$$

Define: To have prior & posterior to have the same parametric form.

Thus, choose such a prior which would make the posterior have the same fn. form

called the conjugate prior.

Since $f(\theta|D)$ depends upon $f(D|\theta)$, priors are decided from the form of $f(D|\theta)$.

Eg 1: Estimate mean of Normal with σ^2 known.

$$f(x|\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

$$f(D|\mu) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

Prior should be normal as well.

$\therefore$ That would make $f(\mu|D)$ Normal.

Let $f(\mu) = N(\mu_0, \sigma_0)$.

Now, $f(\mu|D) = \dfrac{f(D|\mu)\, f(\mu)}{Z}$.

Substituting,

$$f(\mu|D) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right]$$

$$f(\mu|D) \propto \exp\left(-\frac{1}{2}A\right)$$

$$A = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i^2 + \mu^2\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right) - 2\mu\left(\sum_{i=1}^{n}\frac{x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right).$$

Posterior is also Gaussian

suppose $f(\mu|D)$ is $N(\mu_n, \sigma_n)$ then,

$$f(\mu|D) \propto \exp\left[\frac{-1}{2}\left[\frac{\mu^2}{\sigma_n^2} + \frac{\mu_n^2}{\sigma_n^2} - 2\mu\frac{\mu_n}{\sigma_n^2}\right]\right]$$

Comparing,

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

$$\frac{\mu_n}{\sigma_n^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i + \frac{\mu_0^2}{\sigma_0^2}$$

$\Rightarrow$

$$\sigma_n^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \bar{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 .$$

with larger n, μ_0 is initial guess on μ & σ_0 determines the uncertainity.

$$\bar{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{— MLE.}$$

$\mu_n \to$ convex comb of $\bar{\mu}_n$ & μ_0.

Both data & prior contribute.

for $n \to \infty$, $\mu_n \simeq \bar{\mu}_n$

with large n, Bayes $\to$ MLE.

with large n, σ_n becomes very small

uncertainity in μ_0 becomes small.

$$f(x|D) = \int f(x|\mu) f(\mu|D) d\mu$$

$$= \int \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\times \frac{1}{\sigma_n\sqrt{2\pi}} \exp\left(\frac{(\mu-\mu_n)^2}{2\sigma_n^2}\right) d_n$$

Term inside exp can be written as

$$-\left(\frac{x-\mu}{2\sigma^2}\right)^2 - \frac{(\mu-\mu_n)^2}{2\sigma_n^2}$$

$$= -\frac{(\sigma_n^2+\sigma^2)}{2\sigma^2\sigma_n^2}\left[\mu^2 - 2\mu\left(\frac{x\sigma_n^2 + \mu_n\sigma^2}{\sigma^2+\sigma_n^2}\right)\right] - \frac{1}{2}\,\frac{x^2\sigma_n^2 + \sigma^2\mu_n^2}{\sigma^2\sigma_n^2}$$

Completing the square w.r.t μ, leads to a quadratic in x within exp.

(5u)