

Statistical Model

A stats model consists of:

- An identification of RVs of interest
- A specification of (a family of) possible joint distr of the RVs
- Identification of params of those distr that are assumed unknown or observable

Statistical Inference

A procedure that produces a probabilistic statement about parts of statistical model.

Parameter

A characteristic or combination of characteristics that determines the joint distr of RVs of interest.

(e.g. μ, σ^2 for family of normal dists)

Statistic

Suppose observable RVs are X_1, \dots, X_n

Let r be an arbitrary real-valued function of n vars.

RV $T = r(X_1, \dots, X_n)$ is called a statistic.

General Class of Inference Problems

- Prediction: Predict RVs that haven't been observed.
It's called estimation if the unobserved are parameters.
- Statistical Decision:
Given observation, choose a decision from available class of decisions with the property that the consequences of each available decision depend on the unknown value of the param.
It's closely related to hypo testing.
- Experimental Design
Choose the type/amount of experimental data to be collected.

Prior: Distr assigned to parameter θ before making observation.
 Denoted by $\pi(\theta)$

Posterior: Conditional distr of param θ on observed RV $X_1 \dots X_n$.
 Denoted by $\pi(\theta | X_1 \dots X_n)$

- Suppose n RV $X_1 \dots X_n$ form a random sample from distr w/ pdf $f(x|\theta)$, and θ unknown w/ prior $\pi(\theta)$.
 Then posterior of θ :

$$\pi(\theta | \vec{x}) = \frac{\pi(f(x_i|\theta) \cdot \pi(\theta))}{g_n(\vec{x})} \quad \text{for } \theta \in \mathcal{U}$$

$\vec{x} = (x_1 \dots x_n)$ \hookrightarrow marginal joint pdf of $x_1 \dots x_n$

$$g_n(x) = \int_{\mathcal{U}} f_n(\vec{x} | \theta) \pi(\theta) d\theta$$

Note. $g_n(\vec{x})$ is invariant wrt θ ; and thus if view θ as var,
 $\pi(\theta | \vec{x}) \propto f_n(\vec{x} | \theta) \pi(\theta)$

(If RHS has known form, we can normalize w/o integration)

Likelihood Function: Joint pdf $f_n(\vec{x} | \theta)$ of observations given θ .

Posterior interpretation: prior adjusted with likelihood.

Conjugate Prior

Let $X_1 \dots X_n$ be conditionally i.i.d given θ with common pof $f(x|\theta)$. Let Ψ be a family of distr over param space \mathcal{V} .

If: \forall prior $\varepsilon \in \Psi$ we choose, \forall observation $x_1 \dots x_n$,

posterior $\varepsilon(\theta | \vec{x})$ s.t. $\varepsilon(\theta | \vec{x}) \in \Psi$,

Ψ is a conjugate family of prior for samples from $f(x|\theta)$.

• Bernoulli - Beta

Suppose $X_1 \dots X_n$ random sample from Bernoulli w/ param θ , prior of θ is Beta w/ param $\alpha, \beta > 0$,

Then posterior of θ given $x_1 \dots x_n$ is Beta w/

$$\alpha' = \alpha + \sum_{i=1}^n x_i, \quad \beta' = \beta + n - \sum_{i=1}^n x_i$$

• Poisson - Gamma

$X_1, \dots X_n \sim$ Poisson w/ param λ , prior of λ is Gamma

w/ param $\alpha, \beta > 0$, posterior is Gamma w/

$$\alpha' = \alpha + \sum_{i=1}^n x_i, \quad \beta' = \beta + n$$

- Normal - Normal

$X_1 \sim X_n$ random sample from normal w/ unknown mean θ , known var σ^2 . Prior is normal with mean μ_0 , var V_0 .

Posterior is normal w/

$$\mu_1 = \frac{\sigma^2 \mu_0 + n V_0^2 \bar{x}_n}{\sigma^2 + n V_0^2}, \quad V_1^2 = \frac{\sigma^2 V_0}{\sigma^2 + n V_0^2}$$

- Exponential - Gamma

$X_1 \sim X_n$ ~ exponential w/ param θ ,

prior of θ is Gamma w/ $\alpha, \beta > 0$,

posterior is Gamma w/ $\alpha' = \alpha + n$, $\beta' = \beta + \sum_{i=1}^n x_i$

• For Bernoulli related distr, e.g. binomial, geometric
negative binomial, conjugate family for success rate is
beta.

with known first param

• For Poisson related distr, e.g. Poisson, Gamma, exponential
conjugate family of param is gamma.

Estimator / Estimate

Let X_1, \dots, X_n be observable data whose joint distr is indexed by param θ taking values from $\mathcal{S} \subseteq \mathbb{R}$.

An estimator of θ is a real-valued function $f(X_1, \dots, X_n)$.
If $X_1 = x_1, \dots, X_n = x_n$ are observed, $f(x_1, \dots, x_n)$ is called an estimate of θ .

(Estimator is a function, estimate is RV.)

Loss Function

Real-valued func of two var, $L(\theta, a)$, which is loss when param takes value θ while estimate is a .

Consistent Estimator

A sequence of estimators that converges in prob to the unknown value of the parameter being estimated as $n \rightarrow \infty$

Bayes Estimator / Estimate

Given loss func $L(\theta, a)$ for each value \vec{x} of \vec{X} ,

Let $f^*(\vec{x}) = \operatorname{argmin}_a \mathbb{E}[L(\theta, a) | \vec{x}]$,

f^* is a Bayes estimator of θ .

For observed $\vec{X} = \vec{x}$, $f^*(\vec{x})$ is a Bayes estimate.

(f^* will depend on L , prior of θ . It may not exist.)

- For $L(\theta, a) = (\theta - a)^2$, $\mathbb{E}[\theta | \vec{x}] < \infty$

Bayes estimator of θ is $f^*(\vec{x}) = \mathbb{E}[\theta | \vec{x}]$

- For $L(\theta, a) = |\theta - a|$, a Bayes estimator of θ is
 $f^*(\vec{x}) = \text{median of posterior } f(\theta | \vec{x})$

Bayes estimators are consistent under fairly general conditions, and for a wide class of distr.

Specifically, for the distr-prior pair above, and L being square loss, BEs are consistent.

Limitation of Bayes Estimator

It's required to specify a loss function and a prior which is not always available. It's esp. difficult to design a reasonable prior for vector param. Sometimes you are only interested in one dim of the vector param but will still have to design prior for all dim.

MAP (Maximum-a-posteriori) Estimator

(Requires a prior, but not loss function)

Let \vec{x} be a sample of observable RV X .

Given prior of θ , $g(\theta)$, MAP estimator is given by

$$\begin{aligned}\hat{\theta}_{\text{MAP}}(\vec{x}) &= \operatorname{argmax}_{\theta} f(\theta | \vec{x}) \\ &= \operatorname{argmax}_{\theta} \frac{f(\vec{x} | \theta) g(\theta)}{\int_t f(\vec{x} | t) g(t) dt} \\ &= \operatorname{argmax}_{\theta} f(\vec{x} | \theta) g(\theta)\end{aligned}$$

- MAP = ML for uniform prior
- Compared to ML, MAP uses an augmented objective that incorporates prior instead of just likelihood. Therefore it can be seen as a regularization of ML.
- Compared to Bayes estimator, when loss is

$$L(\theta, a) = \begin{cases} 0 & \text{if } |a - \theta| < c \\ 1 & \text{o/w} \end{cases}$$

, as $c \rightarrow 0$, if distribution of θ quasi-concave,

Bayes estimator \rightarrow MAP estimator.

But generally, MAPE is not BE unless θ discrete.

Maximum Likelihood Estimator

A method of choosing estimate that avoids using prior and loss.

It chooses the parameter that maximizes the likelihood function.

MLE Definition

For each observed vector \vec{x} , let $\hat{\theta}(\vec{x}) \in \mathcal{S}$ s.t.

$\hat{\theta}(\vec{x}) = \operatorname{argmax}_{\theta} f(\vec{x}|\theta)$, where f is likelihood function.

$\hat{\theta} = \hat{\theta}(X)$ is ML estimator of θ .

For observed $\vec{X} = \vec{x}$, $\hat{\theta}(\vec{x})$ is an ML estimate.

Notes MLE is required to $\in \mathcal{S}$. There exist cases that $\hat{\theta} \notin \mathcal{S}$.

In those cases we say MLE do not exist.

It's also possible that maximizer $\hat{\theta}$ is not unique. In those cases, any of the maximizer is MLE.

In many practical problem, MLE exists and is unique.

MLE of Important Distr

- Sampling from a Bernoulli distr, $\hat{\theta} = \bar{X}$
- Sampling from a Normal distr
$$\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = \left(\bar{X}, \frac{1}{n} \sum (X_i - \bar{X})^2 \right)$$

sample mean sample var
- Sampling from uniform distr over $[0, \theta]$, θ unknown,
$$\hat{\theta} = \max\{X_1, \dots, X_n\}$$

Limitation of MLE:

- Non-existence
- Non-uniqueness
- Biasness (e.g. when sampling from uniform, $MLE < g + \theta$ almost surely.)

Properties of MLE

- Invariance wrt transformation

If $\hat{\theta}$ is MLE of θ , g is 1-to-1 function,
 $\hat{g} = g(\hat{\theta})$ is MLE of $g(\theta)$.

- Consistency

Given a seq of samples w/ size $1, \dots, n$, MLE seq $\hat{\theta}_n$,
under mild conditions that usually hold for practical problems,
 $\hat{\theta}_n$ converges to the unknown θ .

Method of Moments Estimator (useful when MLE is difficult)

Assume X_1, \dots, X_n random sample from a distr indexed by k-dim param θ .

For $j=1 \dots k$, find function $\mu_j = \mathbb{E}(X_1^j | \theta)$

Let sample moments be $m_j = \frac{1}{n} \sum_i X_i^j$.

To find an estimator of θ , solve θ from $\{m_j = \mu_j\}$

- If first k moments of the distr exist and is finite for $\forall \theta$, and the resulting function μ_j 's are continuous, MDM estimators are consistent.

Expectation Maximization (EM) Estimator

Iterative method to estimate MLE when ML intractable due to unobserved var.

Consider model where X observed var, Z unobserved var, θ unknown param, likelihood $p(X|Z, \theta)$ known.

ML objective: $\max_{\theta} p(X|Z) = \max_{\theta} \int_Z p(X|Z, \theta) p(Z|\theta) dZ$

The integration often causes the optimization to be intractable.

EM iteratively finds an estimate of the problem.

Iterate the following procedure:

In step $j+1$, given former estimation $\theta^{(j)}$,

E step:

- Assume $Z \sim$ conditional distr based on $\theta^{(j)}$: $p(Z|X=\vec{x}, \theta=\theta^{(j)})$
- Compute mean of data likelihood given distr of $Z|\theta$ still treated
(or log likelihood) as var

$$Q(\theta) = \mathbb{E}_{Z \sim p(\cdot|X, \theta^{(j)})} [\log p(x|Z, \theta)]$$

M step: $\theta^{j+1} = \operatorname{argmax}_{\theta} Q(\theta)$