Lecture-14: Point estimation

1 Point estimation

Consider the case when the output space $\mathcal{Y} = \mathbb{R}^d$.

1.1 Bayesian estimation

Consider a parameter family $\Theta \subseteq \mathbb{R}^d$, and parametrized family of probability measures $(P_\theta \in \mathcal{M}(\mathcal{X}) : \theta \in \Theta)$. We assume that for some parameter $\theta \in \Theta$, an unlabeled sample $X \in \mathcal{X}^m$ is generated conditionally *i.i.d.* from the distribution P_θ. We are interested in estimating the parameter θ, under a known prior distribution $\pi \in \mathcal{M}(\Theta)$ on family of parameters. Denoting p_θ as the parametrized density of observation $X : \Omega \to \mathcal{X}^m$, we can write the posterior density of parameter θ given the observation $\{X = x\}$ as

$$p(\theta \mid x) \triangleq \frac{\pi(\theta) p_\theta(x)}{p(x)},$$

where density of observation X is $p(x) \triangleq \int_\Theta d\theta p(x \mid \theta)\pi(\theta)$. We consider the square loss function $L : \Theta \times \Theta \to \mathbb{R}_+$ defined by $L(\theta, \theta') \triangleq \|\theta - \theta'\|^2$ for all parameters $\theta, \theta' \in \Theta$.

Definition 1.1. Bayes estimator of θ with respect to a loss function L is defined as $h : \mathcal{X}^m \to \Theta$ defined for each unlabeled sample $x \in \mathcal{X}^m$ and the following mean taken over random θ generated by posterior distribution $p(\theta \mid x)$,

$$h(x) \triangleq \arg\min_{\theta' \in \Theta} \mathbb{E}[L(\theta, \theta') \mid \{X = x\}].$$

Remark 1. We can re-write the minimization in the right hand side of the definition of Bayes estimator as

$$\arg\min_{\theta' \in \Theta} \mathbb{E}[L(\theta, \theta') \mid \{X = x\}] = \arg\min_{\theta' \in \Theta} \int_\Theta L(\theta, \theta') p_\theta(x) \pi(\theta) d\theta.$$

Example 1.2 (Gaussian mean estimate). Consider an unlabeled sample $X \subseteq \mathcal{X}^m$ where $\mathcal{X} = \mathbb{R}^d$ and unlabeled sample X is *i.i.d.* Gaussian with fixed and unknown mean $\mu \triangleq \mathbb{E}X_1 \in \mathbb{R}^d$ and fixed and known covariance $\Lambda \triangleq \mathbb{E}(X_1 - \mu)(X_1 - \mu)^T \in \mathbb{R}^{d \times d}$. We are interested in estimating the label $y = \mathbb{E}X_1 = \mu$ given sample X. We assume a prior distribution the unknown mean to be Gaussian with zero mean and covariance being identity $\mathbf{I} \in \mathbb{R}^{d \times d}$. for $d = 1$ and $\Lambda = \sigma^2$, we can compute the Bayes estimator for square loss function as

$$\begin{aligned} h(x) &= \arg\min_{y \in \mathbb{R}} \int_{\mathbb{R}} (z - y)^2 \prod_{i=1}^m e^{-\frac{1}{2\sigma^2}(x_i - z)^2} e^{-\frac{1}{2}(z-1)^2} dz \\ &= \arg\min_{y \in \mathbb{R}} \int_{\mathbb{R}} dx (z - y)^2 \exp\Big[-\frac{1 + \frac{m}{\sigma^2}}{2}\Big(z - \frac{(1 + \frac{\sum_{i=1}^m x_i}{\sigma^2})}{(1 + \frac{m}{\sigma^2})}\Big)^2\Big]. \end{aligned}$$

This expression is minimized when $h(x) = \frac{(1 + \frac{\sum_{i=1}^m x_i}{\sigma^2})}{(1 + \frac{m}{\sigma^2})}$, is the mean of examples. For the absolute difference loss function and $d = 1$, it can be shown that the Bayes estimator is median, which is same as mean for Gaussian distribution.

Proposition 1.3. *If the parameter family $\Theta \subseteq \mathbb{R}^d$ is compact then optimal Bayes estimate exists, and if the loss function is strictly convex then the Bayes estimator is unique.*

Remark 2. Under some regularity conditions for sample $x \in \mathcal{X}^m$ with large number of examples, the posterior density $p(\theta \mid x)$ is approximately Gaussian with mean $\theta_0 \in \Theta \subseteq \mathbb{R}^d$ and covariance $I(\theta)^{-1}$, where $I(\theta) = aa^T \in \mathbb{R}^{d\times d}$ is Fisher matrix and $a_i \triangleq \frac{\partial}{\partial \theta_i} p_\theta(x)$ for all $i \in [d]$. This is independent of the prior distribution π, provided the prior distribution is absolutely continuous with respect to Lebesgue measure.

Definition 1.4. Maximum *a posterior* estimator for the parameter θ is defined as

$$h(x) \triangleq \arg\max_{\theta\in\Theta} p(\theta \mid x).$$

Example 1.5 (Gaussian mean estimate). Consider an unlabeled sample $X \subseteq \mathcal{X}^m$ where $\mathcal{X} = \mathbb{R}$ and unlabeled sample X is *i.i.d.* Gaussian with fixed and unknown mean $\mu \triangleq \mathbb{E}X_1 \in \mathbb{R}$ and fixed and known variance σ^2. We are interested in estimating the label $y = \mathbb{E}X_1 = \mu$ given sample X. We assume a prior distribution the unknown mean to be Gaussian with zero mean and unit variance, to write the posterior density

$$p(\mu \mid x) = \frac{1}{\sqrt{\frac{2\pi}{1+\frac{m}{\sigma^2}}}} \exp\Big[-\frac{1+\frac{m}{\sigma^2}}{2}\Big(z - \frac{(1+\frac{\sum_{i=1}^m x_i}{\sigma^2})}{(1+\frac{m}{\sigma^2})}\Big)^2\Big].$$

In this case, MAP estimator and Bayes estimator for square loss functions are identical.

Definition 1.6. An estimator of parameter θ is unbiased if $\mathbb{E}_\theta h(x) = \theta$ for all $\theta \in \Theta$.

Example 1.7 (Gaussian mean estimate). Bayesian estimate of Gaussian examples given by $\hat{\mu}_m \triangleq \frac{1+\frac{1}{\sigma^2}\sum_{i=1}^m x_i}{1+\frac{m}{\sigma^2}}$ is a biased estimate of the mean μ, as $\mathbb{E}\hat{\mu}_m = \frac{\sigma^2+m\mu}{\sigma^2+m}$. When m is large, it becomes an unbiased estimator, since $\lim_{m\to\infty} \mathbb{E}\hat{\mu}_m = \mu$. We also observe that $\lim_{m\to\infty} \hat{\mu}_m = \mu$ almost surely from strong law of large numbers. In addition, it follows from central limit theorem, that $\sqrt{m}(\hat{\mu}_m - \mu)$ converges in distribution to a zero mean normal random variable with variance σ^2.

1.2 Maximum likelihood estimation

Definition 1.8. Maximum likelihood estimator is given by

$$h(x) \triangleq \arg\max_{\theta\in\Theta} p_\theta(x) = \arg\max_{\theta\in\Theta} \log p_\theta(x).$$

Example 1.9 (Gaussian mean estimate). Consider an unlabeled sample $X \subseteq \mathcal{X}^m$ where $\mathcal{X} = \mathbb{R}^d$ and unlabeled sample X is *i.i.d.* Gaussian with fixed and unknown mean $\mu \triangleq \mathbb{E}X_1 \in \mathbb{R}^d$ and fixed and known covariance $\Lambda \triangleq \mathbb{E}(X_1 - \mu)(X_1 - \mu)^T \in \mathbb{R}^{d\times d}$. We are interested in estimating the label $y = \mathbb{E}X_1 = \mu$ given sample X. We can compute the maximum likelihood estimator as

$$h(x) = \arg\min_{y\in\mathbb{R}} \sum_{i=1}^m (x_i - y)^2 = \frac{1}{m}\sum_{i=1}^m x_i.$$

It follows that $h(x)$ is an unbiased estimator of μ. From strong law of larger numbers it follows that $h(x)$ asymptotically converges to μ almost surely in number of examples. From central limit theorem, it follows that $\sqrt{h(x) - \mu} = \frac{1}{\sqrt{m}}\sum_{i=1}^m (x_i - \mu)$ asymptotically converges in distribution to a zero mean Gaussian random variable with variance σ^2.

1.2.1 Asymptotic properties of maximum likelihood estimator

Let $x \in \mathcal{X}^m$ be *i.i.d.* realization from the conditional density p_{θ_0} for some parameter $\theta_0 \in \Theta \subseteq \mathbb{R}^d$.

Proposition 1.10. *Let $\hat{\theta}_m$ be the maximum likelihood estimate of θ_0, then the following are true.*

1. *The shifted and normalized estimate $\sqrt{m}(\hat{\theta}_m - \theta_0)$ converges in distribution to zero-mean Gaussian random variable with covariance $I(\theta)^{-1}$.*
2. *Matrix $I(\theta)^{-1}$ is the minimum covariance.*

Proof. 1. It follows from central limit theorem.

□

Example 1.11 (Gaussian mean estimate). We compute the Fisher information $I(\theta)$ when $p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\theta)^2}$. In this case, we can write

$$A \triangleq \frac{\partial}{\partial\theta} \ln p_\theta(x) = \frac{x-\theta}{\sigma^2}.$$

Therefore, we have $I = \mathbb{E}AA^T = \frac{1}{\sigma^2}\mathbb{E}(X-\mu)^2 = \frac{1}{\sigma^2}$.

2 Machine learning framework

We only have labeled sample $S \in (\mathcal{X} \times \mathcal{Y})^m$. Even if we assume the prior density $\pi \in \mathcal{M}(\mathcal{Y})$, the probability density $p_y(x)$ is unknown. A straightforward approach is to estimate $p_y(x)$ from the sample. However, one may require large number of examples and it maybe computationally challenging for larger number of feature dimensions d, where $\mathcal{X} \subseteq \mathbb{R}^d$.

2.1 Naive Bayes classifier

Assume that features are conditionally *i.i.d.* given label $y \in \mathcal{Y}$. That is, we have $p_y(x) = \prod_{i=1}^d p_y(x_i)$ for any $x \in \mathbb{R}^d$.

Remark 3. One needs to estimate d conditional distributions $p_y(x_i)$ using MLE or Bayes estimator. This estimator outperforms estimating $p_y(x)$.

2.2 Bayes classifier

We assume that $S \in (\mathcal{X} \times \mathcal{Y})^m$ is *i.i.d.* with the common unknown distribution $\mathcal{D}$. We are interested in learning a classifier $h \in \mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ where the set of hypotheses is uncountable, and we assume a prior density π on this set. Using sample S, we obtain a posterior density Q on this set $\mathcal{H}$. Given a loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, the generalization risk for any hypothesis $h \in \mathcal{H}$ is

$$R(h) = \mathbb{E}_{\mathcal{D}}[L(h(X), Y)].$$

If hypothesis h is picked with posterior distribution Q, then the generalization risk is $\mathbb{E}_Q R(h) = \mathbb{E}_Q[L(h(X), Y)]$ and the empirical risk is

$$\mathbb{E}_Q \hat{R}(h) = \frac{1}{m}\sum_{i=1}^m \mathbb{E}_Q[L(h(x_i), y_i)].$$

Theorem 2.1 (PAC Bayes bound). *With probability at least $1-\delta$, we have*

$$\mathbb{E}_Q R(h) \leqslant \mathbb{E}_Q \hat{R}(h) + \sqrt{\frac{D(Q\|\pi) + \ln\frac{m}{\delta}}{2m-1}}.$$

Definition 2.2 (Regularized risk minimization principle). Find the posterior density Q that minimizes the upper bound on the generalization risk, i.e.

$$\arg\min_Q \mathbb{E}_Q \hat{R}(h) + \sqrt{\frac{D(Q\|\pi) + \ln\frac{m}{\delta}}{2m-1}}.$$

Remark 4. A common prior π for hypothesis set $\mathcal{H} = \left\{ \mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| \leqslant \Lambda \right\}$ is the Gaussian distribution.