DENSITY ESTIMATION

Chao Lan

1. Concepts and Notations

Let p_θ be a probability density characterized by a set of unknown parameters θ. Density estimation is the task of estimating θ from a set of observations $S_n = \{x_1, x_2, ..., x_n\}$ that are generated i.i.d. from p_θ.

[Discussion] Given an example of density estimation from an i.i.d. sample.

The estimator $\hat{\theta}$ is a function mapping from S_n to θ. Its output is an estimate of θ, also denoted by $\hat{\theta}$. We will introduce two estimators: maximum likelihood estimation (MLE) and maximum a posteriori (MAP). MLE finds a distribution that can best explain our observations, whereas MAP finds one that can best explain both our observations and beliefs.

2. Maximum Likelihood Estimation (MLE)

MLE finds a θ that maximizes the likelihood of θ on S_n, defined as

$$\ell_n(\theta) = p_\theta(x_1, x_2, ..., x_n) = \Pi_{i=1}^n p_\theta(x_i). \tag{1}$$

[Discussion] Derive (1).

In practice, instead of maximizing the likelihood, we often maximize the log-likelihood, defined as

$$L_n(\theta) = \log \ell_n(\theta) = \log \Pi_{i=1}^n p_\theta(x_i) = \sum_{i=1}^n \log p_\theta(x_i). \tag{2}$$

[Discussion] Derive (2).

[Discussion] Are the solutions of $\max_\theta \ell_n(\theta)$ and $\max_\theta L_n(\theta)$ identical?

[Discussion] Why do we often maximize $L_n(\theta)$ instead of $\ell_n(\theta)$?

2.a. An Example of MLE

Let $x_1, \ldots, x_n$ be a set of instances generated i.i.d. from $\mathcal{N}(\mu, \sigma^2)$. Derive the MLE of μ, assuming σ is known.

Step 0. Recall the probability density function is

$$p_\mu(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}. \tag{3}$$

Step 1. Write the log-likelihood function. Simplify it as necessary.

$$\begin{aligned} L_n(\mu) = \sum_{i=1}^n \log p_\mu(x_i) &= \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\ &= \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi\sigma^2}} + \log \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\ &= \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2, \end{aligned} \tag{4}$$

where the last equality assumes the logarithm function has a natural base.

[Discussion] Derive (4).

Step 2. Maximize $L_n(\mu)$ over μ. Note that it is equivalent to minimizing

$$J(\mu) = \sum_{i=1}^{n}(x_i - \mu)^2. \tag{5}$$

This is a quadratic function of μ. Thus we can solve $J'(\mu) = 0$ for μ and have

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{6}$$

[Discussion] Derive (6).

Note: We can verify $\hat{\theta}$ is indeed a maximum point using the second-order derivative test.

Note: We can also derive the MLE of σ assuming μ is fixed.

2.b. Another Example of MLE

Let $x_1, \ldots, x_n$ be a set of instances generated i.i.d. from $Bernoulli(\theta)$. Derive the MLE of θ.

Step 0. Recall that probability mass function is

$$p_\theta(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}. \tag{7}$$

Step 1. Write the log-likelihood function. Simplify it as necessary.

$$\begin{aligned} L_n(\theta) = \sum_{i=1}^{n} \log p_\theta(x_i) &= \sum_{i=1}^{n} \log \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \sum_{i=1}^{n} x_i \log\theta + (1-x_i)\log(1-\theta) \\ &= \log\theta \sum_{i=1}^{n} x_i + \log(1-\theta)\sum_{i=1}^{n}(1-x_i). \end{aligned} \tag{8}$$

[Discussion] Derive (8).

Step 2. Maximize $L_n(\theta)$ over θ. We can take derivative

$$L_n'(\theta) = \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{1}{1-\theta}\sum_{i=1}^{n}(1-x_i). \tag{9}$$

Setting $L_n'(\theta) = 0$ and solving for θ gives

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{10}$$

[Discussion] Derive (10).

[Discussion] Verify $\hat{\theta}$ is indeed a maximum point using the second-order derivative test.

2.c. A Limitation of MLE

In (10), if we only observe one instance $x_1 = 1$. What would the MLE of θ?

Or, if we observe three instances $x_1 = 1$, $x_2 = 1$ and $x_3 = 1$. What would the MLE of θ?

Are these estimates consistent with your belief on θ ? (e.g., x is the result of flipping a coin)

Can we inject our belief in MLE estimate?

3. Maximum A Posteriori (MAP)

MAP injects our belief into MLE estimate. The belief is modeled as a prior distribution of θ, denoted as $p(\theta)$.

MAP finds a θ that maximizes the posterior of θ, defined as

$$p(\theta; x_1, \ldots, x_n) = \frac{p_\theta(x_1, \ldots, x_n) \cdot p(\theta)}{p(x_1, \ldots, x_n)} \propto \ell_n(\theta)\, p(\theta). \tag{11}$$

[Discussion] Why can we omit the denominator when optimizing over θ?

In practice, instead of maximizing the posterior, one often maximizes the log posterior

$$\log p(\theta; x_1, \ldots, x_n) \propto \log \ell_n(\theta)\; p(\theta) = \log \ell_n(\theta) + \log p(\theta). \tag{12}$$

Note: sometimes it is easier to get posterior first and then take log; sometimes it is easier to take log first.

3.a. An Example of MAP

Let $x_1, \ldots, x_n$ be a set of instances generated i.i.d. from $\mathcal{N}(\mu, \sigma^2)$. Let $p(\mu) \sim \mathcal{N}(0, t^2)$ be a prior of μ. Derive the MAP of μ, assuming σ and t are known.

Step 1. Derive posterior $\log p(\mu; x_1, \ldots, x_n) \propto \log \ell_n(\mu) + \log p(\mu)$. We have $\log \ell_n(\mu)$ in (2.b.). The log prior is

$$\log p(\mu) = -\frac{1}{2}\log(2\pi t^2) - \frac{\mu^2}{2t^2}. \tag{13}$$

Thus

$$\log \ell_n(\mu) + \log p(\mu) = \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{1}{2}\log(2\pi t^2) - \frac{\mu^2}{2t^2}. \tag{14}$$

[Discussion] Derive (13).

Step 2. Find a μ that maximizes the right-hand-side of (14). Note this is equivalent to minimizing

$$J(\mu) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 + \frac{\mu^2}{2t^2}. \tag{15}$$

Setting $J'(\mu) = 0$ and solving for μ gives

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n + \left(\frac{\sigma}{t}\right)^2}. \tag{16}$$

[Discussion] Derive (16).

[Discussion] Compare the MLE and MAP estimates of μ. Assuming σ is fixed. How does t affect the estimate?

3.b. Another Example of MAP

Let $x_1, \ldots, x_n$ be a set of instances generated i.i.d. from $Bernoulli(\theta)$. Assume θ has a prior of a Beta distribution

$$p(\theta; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{17}$$

where α and β are preset constants. We can show the MAP of θ is

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i + \alpha - 1}{n + \alpha + \beta - 2}. \tag{18}$$

[Discussion] Compare the MLE and MAP estimates of θ. If we only observe one instance $x_1 = 1$. What is $\hat{\theta}$ now?

[Discussion] We know Beta is equivalent to uniform when $\alpha = \beta = 1$. How does it affect $\hat{\theta}$?

[Discussion] Can we assume other prior on θ ? (conjugate prior)

4. Summary of MLE and MAP

[Discussion] Summary of MLE and MAP.

5. Learning linear regression models: A density estimation view

Under proper assumptions, we can show that least square and ridge regression for linear model are respectively equivalent to the MLE and MAP estimates of the model in some density estimation tasks.

5.a. MLE = Least Square

Let (x, y) be an observed random instance. Assume y is noisy and is generated by

$$y = x^T\beta + \epsilon, \tag{19}$$

where β is a fixed but unknown parameter, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is noise that does not depend on x.

If x is further fixed, we have

$$y \mid x \sim N(x^T\beta, \sigma^2). \tag{20}$$

[Discussion] Derive (20).

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a set of observations. The log-likelihood of β over the set is

$$\begin{aligned} L_n(\beta) = \sum_{i=1}^{n} \log p_\beta(y_i \mid x_i) &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i^T\beta - y_i)^2\right\} \\ &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^T\beta - y_i)^2. \end{aligned} \tag{21}$$

MLE finds a β that maximizes $L_n(\beta)$. Note this is equivalent to minimizing

$$J(\beta) = \sum_{i=1}^{n}(x_i^T\beta - y_i)^2, \tag{22}$$

which is equivalent to the least square problem.

[Discussions] What if ϵ depends on x? That is, for each x_i there is an $\epsilon_i \sim N(0, \sigma_i^2)$ such that

$$y_i = x_i^T\beta + \epsilon_i. \tag{23}$$

5.b. MAP = Ridge Regression

Continue (5.a.). Let $\tilde{\beta}$ be a vector derived from β by removing its first element (the bias term). Let us assume, as a prior knowledge, that all elements of $\tilde{\beta}$ are generated i.i.d. from $\mathcal{N}(0, t^2)$. This implies $\tilde{\beta}$ has a prior $\mathcal{N}(0, \Sigma)$, where the covariance matrix $\Sigma = t^2 I$ is an identity matrix. The log prior is then

$$\begin{aligned} \log p(\tilde{\beta}) &= \log \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} \exp\left\{-\frac{1}{2}\,\tilde{\beta}^T\Sigma^{-1}\tilde{\beta}\right\} \\ &= \log \frac{1}{\sqrt{(2\pi t^2)^p}} \exp\left\{-\frac{1}{2t^2}\tilde{\beta}^T\tilde{\beta}\right\} \\ &= \log \frac{1}{\sqrt{(2\pi t^2)^p}} - \frac{1}{2t^2}\tilde{\beta}^T\tilde{\beta} \\ &= \log \frac{1}{\sqrt{(2\pi t^2)^p}} - \frac{1}{2t^2}\beta^T I_0 \beta. \end{aligned} \tag{24}$$

[Discussion] Verify (24).

MAP finds a β that maximizes $L_n(\beta) + \log p(\beta)$. Note this is equivalent to minimizing

$$J(\beta) = \sum_{i=1}^{n}(x_i^T\beta - y_i)^2 + \frac{\sigma^2}{t^2} \cdot \beta^T I_0 \beta, \tag{25}$$

which is equivalent to the ridge regression problem with $\lambda = \frac{\sigma^2}{t^2}$ being the regularization coefficient.

[Discussion] Derive (25).

[Discussion] How does σ and t affect the regularization?