Density Estimation Methods

Let $p_\theta$ be a distribution characterized by an unknown parameter $\theta$, and $S_n := x_1, x_2, ..., x_n$ be an observation of $n$ random variables generated from $p_\theta(x)$. Density estimation is the task of estimating $\theta$ from $S_n$ using an estimator $\hat{\theta} : \{S_n\} \to \mathbb{R}$. An output $\hat{\theta}(S_n)$ is an estimate of $\theta$. In this section, we will assume variables are i.i.d. (e.g., the GPA’s of different students are i.i.d.). It is a common in machine learning. It simplifies the designs and analysis of models and the models work well in practice. We will introduce two estimators: maximum likelihood estimation (MLE) and maximum a posteriori (MAP).

**Maximum likelihood estimation (MLE)** finds a $\theta$ that maximizes the likelihood function of $\theta$, which is the joint variable probability

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

where the last equality is based on the i.i.d. assumption.

Theoretically it is often easier to maximize the log-likelihood function

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

[Discussion] Are the maximum points of $\ell_n(\theta)$ and $L_n(\theta)$ identical? Why or why not?

**Example.** Let $x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma^2)$ be a sample of i.i.d. variables. What is the MLE of $\mu$?

Step 1. Let $C_\pi = \log \frac{1}{\sqrt{2\pi \sigma^2}}$. Write down the log likelihood function

$$L_n(\mu) = \sum_{i=1}^n \log p_\theta(x_i) = \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} = \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi \sigma^2}} + \log \exp \left\{ -\frac{(x_i - \mu)^2}{2\sigma^2} \right\} = \sum_{i=1}^n C_\pi - \frac{1}{2\sigma^2} (x_i - \mu)^2 = nC_\pi - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2. \tag{3}$$

Step 2. Find $\mu$ that maximizes $L_n(\mu)$. Here we can apply the critical point method. First,

$$L_n'(\mu) = -\frac{1}{\sigma^2} \sum_{i=1}^n 2(x_i - \mu) = \frac{2}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = \frac{2}{\sigma^2} \left( \sum_{i=1}^n x_i - n\mu \right). \tag{4}$$

Solving $L_n'(\mu) = 0$ for $\mu$ gives

$$\hat{\mu}_{\text{mle}} = \frac{1}{n} \sum_{i=1}^n x_i. \tag{5}$$

$\text{\textsuperscript{1}}$The last equation is based on property $\log AB = \log A + \log B$. 

1
[Exercise] Verify $\hat{\mu}_{mle}$ is the global maximum point (second derivative test + endpoint check).

[Exercise] Derive the MLE of $\sigma$.

MLE suffers from small sample problem. Let $X$ be the random result of a coin flip and $X = 1$ means getting head and $X = 0$ means getting tail. To estimate the probability $\theta$ that $X = 1$ with only one observation $x_1 = 1$, we have $\hat{\theta}_{mle} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{1} = 1$. To address this problem, MAP introduces a prior knowledge on $\theta$, and uses observations to correct the prior.

**Maximum A Posteriori (MAP)** finds a $\theta$ that maximizes the posterior distribution of $\theta$

$$p(\theta; x_1, \ldots, x_n) = \frac{p(y(x_1, \ldots, x_n) \cdot p(\theta)}{p(x_1, \ldots, x_n)} = \ell_n(\theta) \cdot p(\theta),$$

where $p(\theta)$ is a prior distribution of $\theta$ assumed given.

Again, it is often easier to maximize the log posterior

$$\max_{\theta} \log p(\theta; x_1, \ldots, x_n) = \max_{\theta} \log (\ell_n(\theta) \cdot p(\theta)) = \max_{\theta} \ell_n(\theta) + \log p(\theta). \quad (7)$$

**Example.** Let $x_1, \ldots, x_n \sim \mathcal{N}(\mu, \sigma^2)$ be a sample of i.i.d. variables. Let $p(\mu) \sim \mathcal{N}(0, \sigma^2)$ be a prior. What is the MAP estimation of $\mu$?

**Step 1.** Let $C'_\pi = -\frac{1}{2} \log(2\pi \sigma^2)$. Write down the log prior

$$\log p(\mu) = \log \left( \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left\{ -\frac{\mu^2}{2\sigma^2} \right\} \right) = -\frac{1}{2} \log(2\pi \sigma^2) - \frac{\mu^2}{2\sigma^2} = C'_\pi - \frac{\mu^2}{2\sigma^2}. \quad (8)$$

**Step 2.** Write down the log posterior (ignoring the data distribution)

$$\log p(\mu; x_1, \ldots, x_n) \propto \log \ell_n(\theta) + \log p(\theta)$$

$$= nC_\pi - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 + C'_\pi - \frac{\mu^2}{2\sigma^2}$$

$$= nC_\pi + C'_\pi - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{\mu^2}{2\sigma^2}$$

$$= J(\mu). \quad (9)$$

**Step 3.** Find $\mu$ that maximizes $J(\mu)$. Here we can apply the critical point method.

$$J'(\mu) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(x_i - \mu)(-1) - 2\frac{\mu}{2\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) - \frac{1}{\sigma^2} \mu. \quad (10)$$

Solving $J'(\mu) = 0$ for $\mu$ gives

$$\hat{\mu}_{map} = \frac{1}{\frac{1}{\sigma^2} \sum_{i=1}^{n} x_i}{\frac{n}{\sigma^2} + \frac{1}{\sigma^2}} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma^2}} \sum_{i=1}^{n} x_i \quad (11)$$

We can apply the second derivative test to verify that $\hat{\mu}_{map}$ is the maximum point of $J(\mu)$.  

2
Comparing MLE and MAP Estimates

It is interesting to compare the MLE and MAP estimates of $\mu$ in the above two examples.

$$\hat{\mu}_{mle} = \frac{1}{n} \sum_{i=1}^{n} x_i$$
$$\hat{\mu}_{map} = \frac{1}{n + \left(\frac{\sigma_1}{\sigma_2}\right)^2} \sum_{i=1}^{n} x_i.$$  \hspace{1cm} (12)

We see MAP estimate has an additional term $\Delta = (\sigma_1/\sigma_2)^2$.

- If $n$ is big, which means we have sufficient amount of observations, then $\Delta$ can be ignored and $\text{MAP} = \text{MLE}$.

- If $\sigma_2$ is very big, which means $p(\theta)$ is relatively uniform and thus is an uninformative prior, then $\Delta = 0$ and $\text{MAP} = \text{MLE}$.

- If $\sigma_2$ is very small, which means $p(\theta)$ is sharply centered around mean 0 and thus is a strong belief that $\theta = 0$, then $\Delta = \infty$ and $\text{MAP} = 0$ no matter what observation $S_n$ is.

[Discussion] Discuss the impact of $\sigma_1$ on MAP.