Machine Learning, Spring 2019 Chao Lan
EN 2105 EN 4084A
MWF 9-9:50am clan@uwyo.edu

Density Estimation Methods

Let p_θ be a distribution characterized by an unknown parameter θ, 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 θ from S_n using an estimator $\hat{\theta} : \{S_n\} \to \mathbb{R}$. An output $\hat{\theta}(S_n)$ is an estimate of θ. 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 θ that maximizes the likelihood function of θ, which is the joint variable probability

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

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

Theoretically it is often easier to maximize the log-likelihood function[1]

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

[*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 μ?

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

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

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

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

Solving $L_n'(\mu) = 0$ for μ gives

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

[1] The last equation is based on property $\log AB = \log A + \log B$.

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

[*Exercise*] Derive the MLE of σ.

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 θ that $X = 1$ with only one observation $x_1 = 1$, we have $\hat{\theta}_{mle} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{1}1 = 1$. To address this problem, MAP introduces a prior knowledge on θ, and uses observations to correct the prior.

Maximum A Posteriori (MAP) finds a θ that maximizes the posterior distribution of θ

$$p(\theta; x_1, \dots, x_n) = \frac{p_\theta(x_1, \dots, x_n) \cdot p(\theta)}{p(x_1, \dots, x_n)} \propto \ell_n(\theta) \cdot p(\theta), \tag{6}$$

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

Again, it is often easier to maximize the log posterior

$$\max_\theta \log p(\theta; x_1, \dots, x_n) = \max_\theta \log\left(\ell_n(\theta) \cdot p(\theta)\right) = \max_\theta \log \ell_n(\theta) + \log p(\theta). \tag{7}$$

Example. Let $x_1, \dots, x_n \sim \mathcal{N}(\mu, \sigma_1^2)$ be a sample of i.i.d. variables. Let $p(\mu) \sim \mathcal{N}(0, \sigma_2^2)$ be a prior. What is the MAP estimation of μ?

Step 1. Let $C'_\pi = -\frac{1}{2}\log(2\pi\sigma_2^2)$. Write down the log prior

$$\log p(\mu) = \log\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp\left\{-\frac{\mu^2}{2\sigma_2^2}\right\}\right) = -\frac{1}{2}\log(2\pi\sigma_2^2) - \frac{\mu^2}{2\sigma_2^2} = C'_\pi - \frac{\mu^2}{2\sigma_2^2}. \tag{8}$$

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

$$\begin{aligned}
\log p(\mu; x_1, \dots, x_n) &\propto \log \ell_n(\mu) + \log p(\mu) \\
&= nC_\pi - \frac{1}{2\sigma_1^2}\sum\nolimits_{i=1}^{n}(x_i - \mu)^2 + C'_\pi - \frac{\mu^2}{2\sigma_2^2} \\
&= nC_\pi + C'_\pi - \frac{1}{2\sigma_1^2}\sum\nolimits_{i=1}^{n}(x_i - \mu)^2 - \frac{\mu^2}{2\sigma_2^2} \\
&= J(\mu).
\end{aligned} \tag{9}$$

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

$$J'(\mu) = -\frac{1}{2\sigma_1^2}\sum\nolimits_{i=1}^{n} 2(x_i - \mu)(-1) - 2\frac{\mu}{2\sigma_2^2} = \frac{1}{\sigma_1^2}\sum\nolimits_{i=1}^{n}(x_i - \mu) - \frac{1}{\sigma_2^2}\mu. \tag{10}$$

Solving $J'(\mu) = 0$ for μ gives

$$\hat{\mu}_{map} = \frac{\frac{1}{\sigma_1^2}\sum_{i=1}^{n} x_i}{\frac{n}{\sigma_1^2} + \frac{1}{\sigma_2^2}} = \frac{1}{n + \left(\frac{\sigma_1}{\sigma_2}\right)^2}\sum\nolimits_{i=1}^{n} x_i \tag{11}$$

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

Comparing MLE and MAP Estimates

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

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

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

- For fixed σ_1, σ_2, if n is big (large sample), then Δ becomes negligible and MAP $\to$ MLE.

- For fixed σ_1, if σ_2 is big (μ is uniformly distributed), then $\Delta = 0$ and MAP $\to$ MLE.

- For fixed σ_1, if σ_2 is small (strong belief $\mu = 0$), then $\Delta = \infty$ and MAP $= 0$ disregarding x_i.

[*Discussion*] Discuss the impact of σ_1 on MAP.