CS5487 Problem Set 5

Non-parametric estimation and clustering

Antoni Chan
Department of Computer Science
City University of Hong Kong

Kernel density estimators

Problem 5.1 Bias and variance of the kernel density estimator

In this problem, we will derive the bias and variance of the kernel density *estimator*. Let $X = \{x_1, \cdots, x_n\}$ be the r.v. samples, drawn independently according to the true density $p(x)$.

(a) Show that the mean of the estimator is

$$\mathbb{E}_X[\hat{p}(x)] = \int p(\mu)\tilde{k}(x-\mu)d\mu = p(x) * \tilde{k}(x), \tag{5.1}$$

$E_X[\frac{1}{n}\sum \tilde{k}(x-x_i)]$

$= \frac{1}{n} \times n \cdot E_z[\tilde{k}(x-z)] = E_z[\tilde{k}(x-z)]$

$= \int p(z) \cdot \tilde{k}(x-z) \cdot dz$

where $*$ is the convolution operator. What does this tell you about how the KDE is biased?

(b) Show that the variance of the estimator is bounded by

$$\text{var}_X(\hat{p}(x)) \leq \frac{1}{nh^d} \max_x(k(x))\mathbb{E}[\hat{p}(x)]. \tag{5.2}$$

$E_X[\hat{p}(x)^2] - (E_X[\hat{p}(x)])^2$

$\leq E_X[\hat{p}(x)^2]$

$= \frac{1}{n^2} \times n \, E_z[\tilde{k}(x-z)^2]$

$= \frac{1}{n}\int \tilde{k}(x-z)^2 p(z) \cdot dz$

$\leq \frac{1}{n}\tilde{k}(x^*) \, E[\hat{p}(x)]$

$= \frac{1}{nh^d} \cdot k(x^*) \, E[\hat{p}(x)]$

Hint: the following properties will be helpful:

$$\text{var}(x) = \mathbb{E}[x^2] - (\mathbb{E}[x])^2 \leq \mathbb{E}[x^2], \tag{5.3}$$

$$k\left(\frac{x-x_i}{h}\right) \leq \max_x k(x), \tag{5.4}$$

and Problem 1.4.

. . . .

Problem 5.2 Mean and variance of a kernel density estimate

In this problem, we will study the mean and variance of the kernel density *estimate*, i.e., the distribution $\hat{p}(x)$. Let $X = \{x_1, \cdots, x_n\}$ be the set of samples, and $\tilde{k}(x)$ be the kernel with bandwidth included. The estimated probability distribution is

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \tilde{k}(x - x_i). \tag{5.5}$$

Suppose that the kernel function $\tilde{k}(x)$ has zero mean and covariance H, i.e.,

$$\mathbb{E}_{\tilde{k}}[x] = \int \tilde{k}(x)xdx = 0, \tag{5.6}$$

$$\text{cov}_{\tilde{k}}(x) = \int \tilde{k}(x)(x - \mathbb{E}_{\tilde{k}}[x])(x - \mathbb{E}_{\tilde{k}}[x])^T dx = H. \tag{5.7}$$

$$\int \frac{1}{n}\sum_i \tilde{k}(x-x_i)\cdot x\cdot dx = \frac{1}{n}\int \sum_i \tilde{k}(x-x_i)\, x\, dx = \sum_i \frac{1}{n}\int \tilde{k}(x-x_i)\cdot(x-x_i)\, d(x-x_i) + \sum_i \frac{1}{n}\int \tilde{k}(x-x_i)\cdot x_i\, dx = \frac{\sum_i x_i}{n}$$

(a) Show that the mean of the distribution $\hat{p}(x)$ is the sample mean of X,

$$\hat{\mu} = \mathbb{E}_{\hat{p}}[x] = \int \hat{p}(x)x dx = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{5.8}$$

(b) Show that the covariance of the distribution $\hat{p}(x)$ is

$$\hat{\Sigma} = \int \frac{1}{n}\sum_i \tilde{k}(x-x_i)(x-\bar{\mu})(x-\bar{\mu})^T dx$$

similarly substract $x-x_i$

$$\hat{\Sigma} = \text{cov}_{\hat{p}}(x) = H + \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})^T, \tag{5.9}$$

where the second term on the right hand side is the sample covariance.

(c) What does this tell you about the properties of the kernel density estimate $\hat{p}(x)$? How does this relate to the bias of the kernel density estimator?

If the estimator is unbiased, $\Rightarrow$ $\hat{\mu}$ and $\hat{\Sigma}$ will be the mean/cov of the data.

....

2 points of view

5.2a\b算出来的结论和bias of kd estimator的关系就是
这个bias是由模糊导致的，而模糊不会改变均值只会增加方差（毕竟更分散了）h影响H

Problem 5.3 KDE with Gaussians

Consider the kernel function $k(x) = \mathcal{N}(x|0,1)$, and samples $X = \{x_1,\ldots,x_n\}$ generated from a Gaussian, $p(x) = \mathcal{N}(x|\mu,\sigma^2)$. Show that the kernel density estimate,

$X \Rightarrow$ data (x_i) expected over data. $\tilde{k}(x) = \mathcal{N}(x|0,d^2)$

$$\hat{p}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} k\left(\frac{x-x_i}{h}\right), \tag{5.10}$$

has the following properties, for small h:

x is fixed since after $\mathbb{E} \Rightarrow x$ is variable.

(a) $\mathbb{E}_X[\hat{p}(x)] = \mathcal{N}(x|\mu,\sigma^2+h^2)$.

(b) $\text{var}_X(\hat{p}(x)) \approx \frac{1}{2nh\sqrt{\pi}}p(x)$.

(c) $\text{bias}(\hat{p}(x)) = p(x) - \mathbb{E}_X[\hat{p}(x)] \approx \frac{h^2}{2\sigma^2}\left[1 - \frac{(x-\mu)^2}{\sigma^2}\right]p(x)$.

(d) Setting h as a function of n, $h = a/\sqrt{n}$, what is the convergence rate of the bias and variance of the estimator, in terms of the number of samples n? How does the convergence rate compare with that of the ML estimator for a Gaussian?

....

Problem 5.4 KDE with exponential kernel

Let the true density $p(x) \sim U(0,a)$ be a uniform density from 0 to a. Let the kernel function be

$$k(x) = \begin{cases} e^{-x}, & x > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{5.11}$$

(a) Show that the mean of the kernel density estimator is

$$\mathbb{E}[\hat{p}(x)] = \begin{cases} 0, & x < 0 \\ \frac{1}{a}(1-e^{-x/h}), & 0 \le x \le a \\ \frac{1}{a}(e^{a/h}-1)e^{-x/h}, & a \le x. \end{cases} \tag{5.12}$$

(b) Plot $\mathbb{E}[\hat{p}(x)]$ versus x for $a = 1$ and $h = \{1, \frac{1}{4}, \frac{1}{16}\}$.

(c) How small does h need to be to have less than 1% bias over 99% of the range $0 < x < a$?

(d) Find h for this condition if $a = 1$, and plot $\mathbb{E}[\hat{p}(x)]$ in the range $0 \leq x \leq 0.05$.

....

Mean-shift algorithm

Problem 5.5 Epanechnikov kernel

Consider the Epanechnikov kernel,

$$k(x) = \begin{cases} \frac{d+2}{2c_d}(1 - \|x\|^2), & \|x\|^2 < 1 \\ 0, & \text{otherwise}, \end{cases} \tag{5.13}$$

where c_d is the volume of a d-dimensional sphere.

(a) What is the kernel profile $\bar{k}(r)$ of the Epanechnikov kernel? Will the mean-shift algorithm converge when this kernel is used?

(b) What is the corresponding kernel profile $\bar{g}(r)$?

(c) Write down the mean-shift updates of the mode $\hat{x}^{(k+1)}$ using the Epanechnikov kernel. Compare these updates with the mean-shift updates using the Gaussian kernel.

(d) Comment on the similarities/differences between the mean-shift algorithm using the Epanechnikov or Gaussian kernels, the K-means algorithm, and EM for GMMs?

....

Problem 5.6 Finding local modes in a GMM

In this problem, you will consider using mean-shift to find the local modes (peaks) in a Gaussian mixture model. Note that the mean of a Gaussian component does not always correspond to a peak of the GMM, since the other components can influence the peak location.

(a) Derive an algorithm similar to mean-shift to find the modes of a Gaussian mixture model,

$$p(x) = \sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j). \tag{5.14}$$

(b) Will this algorithm converge to a stationary point? Why or why not?

(c) Suppose we choose the means $\{\mu_j\}_{j=1}^{K}$ as the starting points of the algorithm. In this case, will it find all the modes of the GMM? If not, can you construct a counter example?

In D>=2, even a mixture of isotropic
Gaussians can have more than N modes
(Carreira-Perpiñán & Williams ’03):

....