Chapter 8: Unsupervised Learning -- Density Estimation

- Recap of Basics of Probability Theory
  - Probability Density Function
  - Some well-known Probability Distributions

- Recall that: Unsupervised learning works with datasets that do not have labels.
  - Its main target is to understand the structure of the data.
- In many scenarios,
  - the observations are generated randomly
  - following a probability distribution.
    e.g. Gaussian/normal distribution.

Example: If you want to generate 200 real numbers which following the Gaussian distribution with mean = 6.4 and variance 12, the R code is

```
rnorm(200, 6.4, sqrt(12))
```

Problem: - You don't know what the probability distribution is;
- you only have access to the observations.

Density Estimation: - A learning process which estimates the probability distribution,
- by using the observations we have.

Probability Density Function

- When we talk about generating random samples, the outputs are either
  1. categorical / discrete
     e.g. apple / orange / banana; throw a dice
  2. quantitative / continuous
     e.g. sample a real number in the interval $[0, 1]$

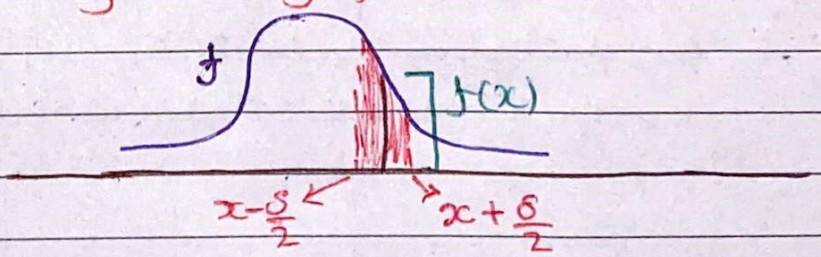
- For the discrete case,
  - we specify the distribution for each possible output:

$$P[X = \text{apple}] = \frac{1}{2} \quad P[X = \text{orange}] = \frac{1}{3} \quad P[X = \text{banana}] = \frac{1}{6}$$

$$P[X = i] = \frac{1}{6} \text{ for every } i = 1, 2, \ldots, 6$$

  - Each probability is a real number between 0 and 1, and the total sum is 1.

- For the continuous case,
  - there are infinitely many possible outputs.
  - we specify the distribution via a probability density function.



Intuitively, given a probability density function f,
- at any $x \in \mathbb{R}$ and for any small $\delta > 0$,

$$P\left[x - \frac{\delta}{2} \leq X \leq x + \frac{\delta}{2}\right] = \text{area of the pink region} \approx f(x) \cdot \delta$$
$$= f(x) \cdot (\text{width})$$

- There are analogous constraints for p.d.f. f:
  a) $f(x) \geq 0$ for all $x \in \mathbb{R}$
  b) $\int_{-\infty}^{+\infty} f(x)\,dx = 1$ (i.e., the total area under the p.d.f. is 1)

Well-known Probability Density Functions

1. Uniform distribution over the interval $[a, b]$:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

2. Gaussian distribution with mean μ and variance σ^2:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x-\mu)^2}{(2\sigma^2)}}$$

3. Exponential distribution with parameter θ:

$$f(x) = \begin{cases} \theta \cdot e^{-\theta x} & \text{if } x \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

- All these p.d.f. can be specified by a few parameters, -so they are sometimes called parametrized distributions

Example:

- The uniform distribution is specified by the parameters a, b.
- The Gaussian distribution is specified by the parameters μ, σ^2.

Parametric Density Estimation

- Maximum Likelihood Principle
- Application to Different Families of p.d.f.

- Recall that: Our target is to estimate the underlying probability distribution.
- Sometimes you know/assume the type of parametrized distribution
  e.g. - The observations are coming from a Gaussian distribution
  - with unknown parameters of mean and variance.
- One standard approach is to choose the parameter that maximize likelihood. (Recall logistic regression)

Maximum likelihood principle

Let the dataset be $\{x_1, x_2, \ldots, x_n\}$

1. Let $\theta(\theta_1, \theta_2, \ldots, \theta_p)$
   - be the parameters of the parametrized distribution.
   - let f_θ denote the distribution's probability density function.

2. For the i^{th} observation,
   - $f_\theta(x_i)$ is the likelihood that x_i is sampled under the parameters θ.

3. Assuming that each observation is generated independently,
   - then the likelihood that the dataset is sampled
   - is the product of individual likelihoods:

$$f_\theta(x_1) \times f_\theta(x_2) \times \ldots \times f_\theta(x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

4. compute θ that maximizes the above product

5. Usually, it is easier to maximize the logarithm of the product.

i.e. maximize $\sum_{i=1}^{n} \log f_\theta(x_i)$

- we call this sum the log-likelihood.

Theorem. If the distribution is assumed to be Gaussian,
- then the parameters choice

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

- maximizes the log-likelihood.

Proof: we want to maximize the log-likelihood for the Gaussian case.

- The log-likelihood is

$$g(\mu,\sigma^2) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)\right)$$

$$g(\mu,\sigma^2) = -n\log\left(\sqrt{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

- At maximum $\frac{\partial g}{\partial \mu} = \frac{\partial g}{\partial \sigma^2} = 0$

- By simple calculus,

$$\frac{\partial g}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(\mu - x_i) \quad \text{and} \quad \frac{\partial g}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(\mu-x_i)^2$$

- Solving $\frac{\partial g}{\partial \mu} = 0$ and $\frac{\partial g}{\partial \sigma^2} = 0$

Theorem: If the distribution is assumed to be Exponential,
- then the parameter choice

$$\theta = 1/\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

- maximizes the log-likelihood

Proof: We want to maximize the log-likelihood for Exponential case.
- The log-likelihood is

$$g(\theta) = \sum_{i=1}^{n} \log\left(\theta \cdot e^{-\theta x_i}\right) = n\log\theta - \theta\sum_{i=1}^{n} x_i$$

- At maximum, $\frac{\partial g}{\partial \theta} = 0$
- By simple calculus, $\frac{\partial g}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} x_i$
- Solving $\frac{\partial g}{\partial \theta} = 0$ so, $\frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0$

Summary of Parametric Density Estimation.

- Parametrized Density Estimation
  - assumes the underlying distribution
  - belongs to a specific family of parametrized distributions.
  - This simplifies the problem to computing the "optimal" parameters.
- One standard approach to define "optimality" is via
  - the maximum likelihood principle
  - which means computing the parameters which maximize the log-likelihood.

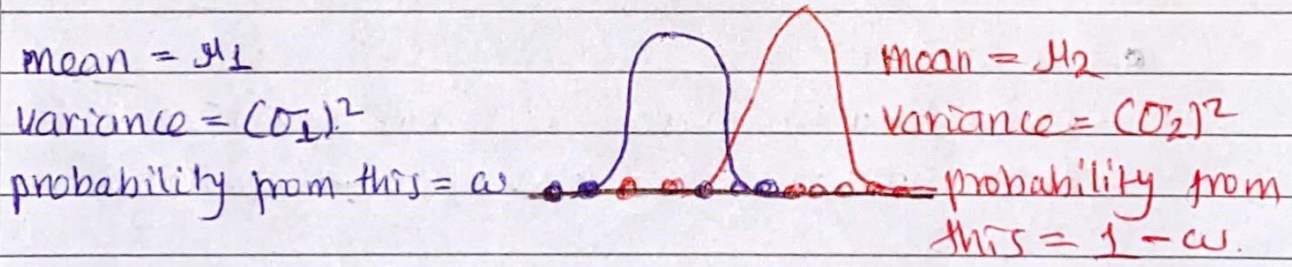
Mixtures of Gaussians

- model
- Expectation-Maximization (EM) algorithm

- Now we turn to a natural but more involved
  - model of parametric density estimation
- Suppose the observations in a dataset are coming from different random process,
  e.g. - the observations are coming from a few
  - distinct Gaussian distribution with different means and variances.

For example

- the observations can be the polling data
- but due to some limitations
- the data collector is unable to identify the age, gender, race, income of the people they interviewed.



- If we know from which Gaussian each observation is generated,
  - then we can use maximum likelihood principle
  - for each group to estimate the two Gaussians.

But the challenge is: you do not really know the colors/classes...

- There are K different Gaussian distributions;
  - The j^{th} one is chosen with probability w_j,
  - and it has mean μ_j and variance $(\sigma_j)^2$
  - we need $\sum_{j=1}^{K} w_j = 1$

- Thus, the p.d.f. of the j^{th} Gaussian is

$$f_j(x) = \frac{1}{\sqrt{2\pi(\sigma_j)^2}} \cdot e^{-\frac{(x-\mu_j)^2}{(2(\sigma_j)^2)}}$$

- The overall probability density function is

$$f(x) = \sum_{j=1}^{K} w_j \cdot f_j(x)$$

Note that: now the parameters are

$$\mu_1, (\sigma_1)^2, w_1, \ldots, \mu_K, (\sigma_K)^2, w_K$$

- Our target
  - is to find the parameters that maximize log-likelihood,
  - or minimize the negative of log-likelihood.

- The issue is, we do not have exact formula now,
  - so we ought to use methods like gradient descent (G.D).
- Another popular algorithm is
  - Expectation-Maximization (EM) algorithm.

- Both G.D and EM only converge to local minimum.

Expectation-Maximization (EM) Framework

- **Model:**
  - There are k different Gaussian distributions;
  - the j^{th} one is chosen with probability w_j,
  - and it has mean μ_j and variance $(\sigma_j)^2$

- The observations in the dataset $\{x_1, x_2, \ldots, x_n\}$,
  - where each $x_i \in \mathbb{R}$
- To fit into the EM framework,
  - we view each observation x_i as (x_i, y_i)
  - where $y_i \in \{1, 2, \ldots, k\}$ is a hidden label.

- The EM algorithm proceeds by repeating E-step & M-step

E-step: use the parameters to estimate the hidden labels $y_1, y_2, \ldots, y_n$

Parameters of the model $(w_j, \mu_j, (\sigma_j)^2)$ → "completion" of the hidden data.

"completion" of the hidden data → Parameters of the model $(w_j, \mu_j, (\sigma_j)^2)$

M-step: - use the "completed" data to update the parameters
  - the updates are motivated by
  - maximum likelihood principle for one Gaussian

E-Step

- If an observation x_i is generated by the j^{th} Gaussian,
- the probability density is $f_j(x_i)$
- which depends on the parameter $\mu_j, (\sigma_j)^2$.

- In probability theory, this can be done via Bayes theorem:

$$\gamma_j(x_i) := P[j|x_i] = \frac{P[j] \cdot P[x_i|j]}{P[x_i]} = \frac{w_j \cdot f_j(x_i)}{\sum_{j=1}^{k} w_j \cdot f_j(x_i)}$$

M-Step

- To understand what happens in M-step,
- you should recall what happens in parametric density estimation for one Gaussian.

• Updates:

$$n_j \leftarrow \sum_{i=1}^{N} \gamma_j(x_i) \qquad w_j \leftarrow \frac{n_j}{N}$$

$$\mu_j \leftarrow \frac{1}{n_j} \sum_{i=1}^{N} \gamma_j(x_i) \cdot x_i$$

$$(\sigma_j)^2 \leftarrow \frac{1}{n_j} \sum_{i=1}^{N} \gamma_j(x_i) \cdot (x_i - \mu_j)^2$$

EM for Mixtures of Gaussians

1. Choose the initial parameter value of $w_j, \mu_j, (\sigma_j)^2$ for $j = 1, 2, \ldots, K$.

2. Repeat the E-step & M-step
   - below for multiple times, until convergence:

• E-step: for each observation x_i and each $j = 1, 2, \ldots, K$, compute

$$\gamma_j(x_i) := P[j|x_i] = \frac{w_j \cdot f_j(x_i)}{\sum_{j=1}^{K} w_j \cdot f_j(x_i)}$$

• M-step: for $j = 1, 2, \ldots, K$, update the parameters as below:

$$n_j \leftarrow \sum_{i=1}^{N} \gamma_j(x_i) \qquad w_j \leftarrow n_j/N \qquad \mu_j \leftarrow \frac{1}{n_j} \sum_{i=1}^{N} \gamma_j(x_i) \cdot x_i$$

$$(\sigma_j)^2 \leftarrow \frac{1}{n_j} \sum_{i=1}^{N} \gamma_j(x_i) \cdot (x_i - \mu_j)^2$$

Summary of Parametric Density Estimation

- Datasets often contain observations from different random processes.
- We are interested in estimating all these random processes.
- In the special case that each random process follows a Gaussian distribution, the codename is "Mixture of Gaussians".

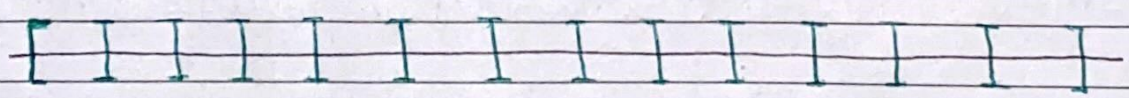
- Although maximum-likelihood principle plus gradient descent can be used for "Mixture of Gaussians"
- a more popular and easy-to-implement algorithm is the Expectation-Maximization (EM) algorithm

Non-Parametric Density Estimation

- Histogram and Curse of Dimensionality
- Kernel Density Estimation.

Histogram

- The first method in non-parametric density estimation



Suppose there are n observations.

- Cut the set of real numbers into many small disjoint "bins".
- For each bin, see how many observations fall into it.
- Then our estimate is, for each bin,

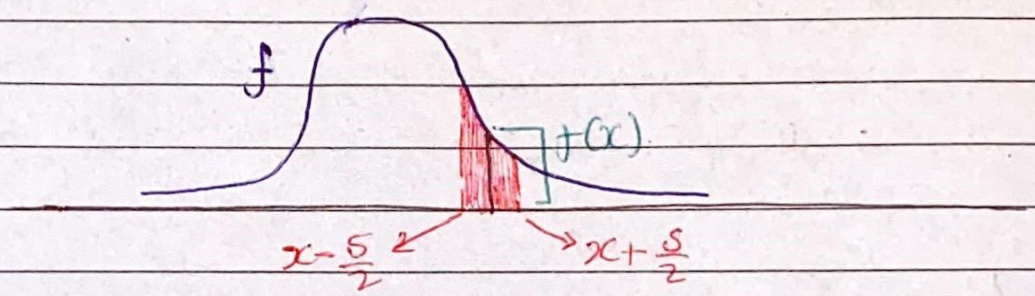
$$P[X \text{ lies in the bin}] = \frac{(\text{number of observations in the bin})}{n}$$

- Hence, our estimate of the density at any point x in the bin is

$$f(x) = \frac{1}{(\text{length of the bin})} \cdot \frac{(\text{number of observations in the bin})}{n}$$

- To be accurate, (n° of observation) >> (n° of bins)
- Intuitively, given a probability density function f, at any $x \in \mathbb{R}$ and for any small $\delta > 0$,

$$P\left[x - \frac{\delta}{2} \leq X \leq x + \frac{\delta}{2}\right] = \text{area of the pink region} \simeq f(x) \cdot \delta$$
$$= f(x) \cdot (\text{width})$$



Curse of Dimensionality

- For the histogram method to be accurate, we need:
  (number of observations) >> (number of bins)

- It works nicely when the observations are 1-D ($x_i \in \mathbb{R}$)
- However, if the observations are multi-dimensional...

- If the observations are in $\mathbb{R}^d$,
  - then the number of bins is c^d,
    where, c is the number of bins along each dimension.
  - when d is large, the number of bins increases quickly.

Example:

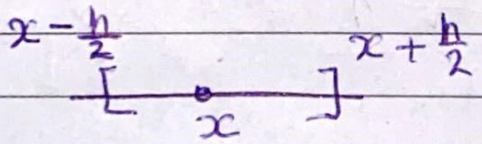
Suppose $c = 10$. When $d = 2$, $c^d = 10^2 = 100$.
When $d = 6$, $c^d = 10^6 = 1,000,000$

So, in this case you need much more than one million observations...

- The curse of dimensionality
  - indicates that histogram is not a good method for high-dimensional data.
- Recall that: the number of bins is c^d.
  We are facing two contrasting agendas of choosing c:
  - we want c to be large enough,
    - so that the grid of bins is fine enough to provide accurate estimates;
  - we want c to be small enough,
    - so that the number of bins (and hence the number of observations needed) is small.
- This leads us to kernel density estimation.

Kernel Density Estimation

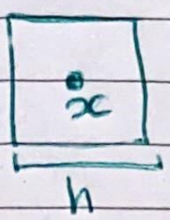
- Suppose the side length of each bin is h,
  - which is called the bandwidth.
  - when h is smaller, there are more bins.



- We define a kernel function K on vector $u \in \mathbb{R}^d$

$$K(u) = \begin{cases} 1 & \text{if } |u_j| \leq \frac{1}{2} \text{ for all } j = 1, 2, \ldots, d; \\ 0 & \text{otherwise.} \end{cases}$$

- This kernel corresponds to a unit hypercube centered at the origin, and is known as the Parzen window.



- For any $x \in \mathbb{R}$, the quantity $K\left(\frac{x - x_i}{h}\right) = ?$
  - If x_i lies inside a hypercube of side length h centered at x,
  - and it is 0 otherwise.

$$K\left(\frac{x - x_i}{h}\right) = \begin{cases} 1 & \text{if } x_i \text{ lies inside a hypercube of side length } h \text{ centered at } x. \\ 0 & \text{otherwise} \end{cases}$$

- Thus, $\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$
  - is the number of observations which lie inside the the hypercube.

- Recall: the formula we derived when discussing histogram:

$$f(x) = \frac{1}{(\text{volume of the bin})} \cdot \frac{(\text{number of observations in the bin})}{n}$$

$$f(x) = \frac{1}{h^d} \cdot \frac{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}{n} = \frac{1}{nh^d} \cdot \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

  - h^d is the volume of each bin.
  - Estimating $f(x)$ for any $x \in \mathbb{R}^d$.

Kernel Density Estimation Steps

1. Suppose the side length of each bin is h, which is called the bandwidth.
2. Define a kernel function K on vector $U \in \mathbb{R}^d$ which is called the Parzen window:

$$K(U) = \begin{cases} 1 & \text{if } |U_j| \leq \frac{1}{2} \text{ for all } j = 1, 2, \ldots, d; \\ 0 & \text{otherwise.} \end{cases}$$

3. The density estimation at any $x \in \mathbb{R}^d$ is

$$f(x) = \frac{1}{nh^d} \cdot \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Note that: this density estimation depends on the choice of h.
- we will see how this choice affects the outcome.

Example:

- Kernel Density Estimation (KDE) works for any dimension,
  - but for illumination we focus on 1D data.
  - The the formula becomes

$$K(U) = \begin{cases} 1 & \text{if } |U| \leq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \qquad f(x) = \frac{1}{nh} \cdot \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Example: Given the dataset $D = \{3, 4, 4, 4, 4, 6, 7, 7, 8, 8, 8, 11\}$. Use the Parzen window with bandwidth $h = 4$ to estimate the density at $x = 5$.

- 12 observation $\in \mathbb{R}$
- $f(x) = \frac{1}{12 \times 4} \sum_{i=1}^{n} K\left(\frac{5 - x_i}{4}\right)$

$$f(5) = \frac{1}{48}\left[K\left(\frac{5-3}{4}\right) + 4 \cdot K\left(\frac{5-4}{4}\right) + K\left(\frac{5-6}{4}\right) + 2K\left(\frac{5-7}{4}\right) + 3K\left(\frac{5-8}{4}\right) + K\left(\frac{5-11}{4}\right)\right].$$

$$f(5) = \frac{1}{48}\left[K\left(\frac{2}{4}\right) + 4K\left(\frac{1}{4}\right) + K\left(\frac{-1}{4}\right) + 2K\left(\frac{-2}{4}\right) + 3K\left(\frac{-3}{4}\right) + K\left(\frac{-6}{4}\right)\right]$$

c. $K(u) = \begin{cases} 1 & \text{if } |u| \le \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$

$$f(5) = \frac{1}{48}\left[1 + 4\times 1 + 1 + 2\times 1 + 3\times 0 + 0\right].$$

$$f(5) = \frac{1}{48}\left[1+4+1+2\right] = \frac{1}{48}\left[8\right] = 1/6 = 0.1667.$$

Quiz ② Given the dataset $D = \{3, 4, 4, 4, 4, 6, 7, 7, 8, 8, 8, 11\}$,
- use the parzen window with bandwidth $h = 3$
- to estimate the density at $x = 7.5$

$$f(7.5) = \frac{1}{12\times 3} \cdot \sum_{i=1}^{n} K\left(\frac{7.5 - x_i}{3}\right)$$

$$= \frac{1}{36}\left[K\left(\frac{7.5-3}{3}\right) + 4K\left(\frac{7.5-4}{3}\right) + K\left(\frac{7.5-6}{3}\right) + 2K\left(\frac{7.5-7}{3}\right) + 3K\left(\frac{7.5-8}{3}\right) + K\left(\frac{7.5-11}{3}\right)\right]$$

$$= \frac{1}{36}\left[K\left(\frac{4.5}{3}\right) + 4K\left(\frac{3.5}{3}\right) + K\left(\frac{1.5}{3}\right) + 2K\left(\frac{0.5}{3}\right) + 3K\left(\frac{-0.5}{3}\right) + K\left(\frac{-3.5}{3}\right)\right]$$

$$= \frac{1}{36}\left[0 + 0 + 1 + 2\times 1 + 3\times 1 + 0\right] = \frac{1}{36}\left[1+2+3\right] = \frac{1}{36}\times 6 = 0.1667$$

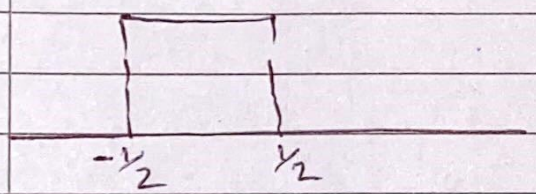
- We plot the probability density function estimated
  - via KDE with Parzen windows,

- Parzen windows, while being naturally motivated by histogram,
  - has certain drawbacks
1. It yields estimates that have discontinuities.
2. within each window, it weights all points x_i equally, regardless of their distances to the estimation point x.

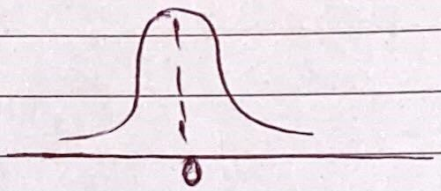
- **Parzen window**
  - is commonly replaced with a smooth kernel function K
  - such that $$\int_{-\infty}^{\infty} K(x)\,dx = 1$$
  - we also require that $$\int_{-\infty}^{\infty} x \cdot K(x)\,dx = 0.$$
  - For instance, we may replace K by the standard Gaussian distribution:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \qquad f(x) = \frac{1}{nh} \cdot \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Parzen window



Gaussian window



Summary of Kernel Density Estimation (KDE)

- KDE incorporates a flexibility to choose the bandwidth h,
  - which corresponds to the side length of bin in histogram.
- Parzen window is naturally motivated
  - but has some drawbacks,
  - so smoother kernel functions are used in practice.
- Choosing h is crucial in KDE:
  - when h is small, there are more likely spikes
  - and thus the probability density function is less smooth.
  - In other words, variance is large.

- when h is large, the probability density function is smooth,
  - but perhaps too smooth to mask out the structure of the data
  - In other words, bias is large.

- There is bias-variance trade-off