if we do not know $p(x)$ (actually, $p^*(x)$), but we have $TS = \{X_i \sim p^*(x)\}_{i=1}^{N}$

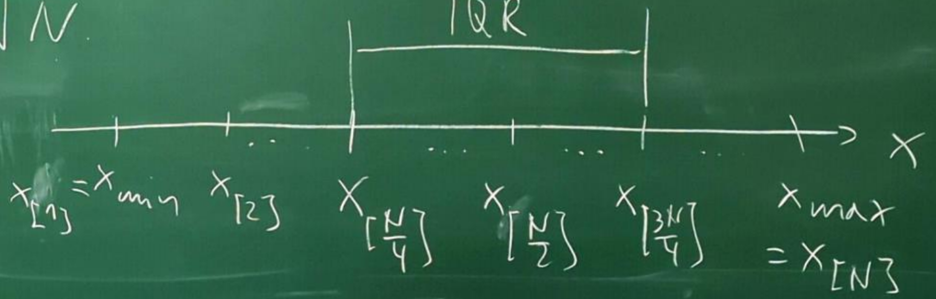
$\Rightarrow$ "fit" our $p(x)$ to the data

(1) histogram   define $X_{min} = \min_i X_i$ smallest value

$$h = \frac{2\, IQR(TS)}{\sqrt[3]{N}} \quad \text{bin width}$$

$IRQ(TS)$: "inter-quartile range"

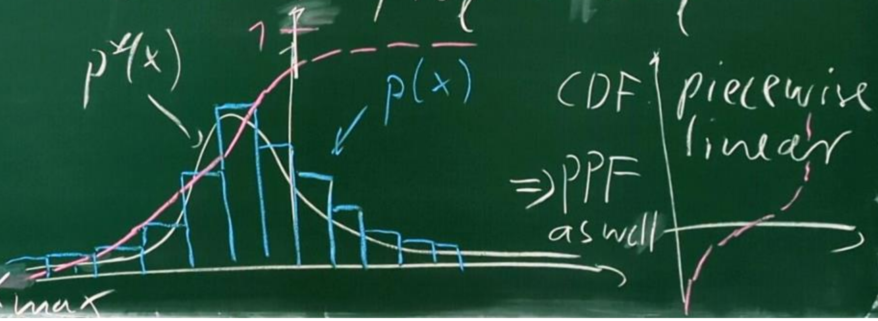
$IRQ(TS) = X_{\left[\frac{3N}{4}\right]} - X_{\left[\frac{N}{4}\right]}$ ← sorted order



$bin_\ell = \{X_i : X_{min} + (\ell-1)h \le X_i < x_{min} + \ell \cdot h\}$, $N_\ell = \#\, bin_\ell$

$$\boxed{p(x) = \sum_{\ell=1}^{L} \mathbb{1}[x \in bin_\ell] \cdot \frac{N_\ell}{N \cdot h}}$$

L: total # bins, such $x_{min} + L \cdot h \ge x_{max}$



generalization: mixture distribution

idea: express complicated $p(x)$ as a superposition of L simple $p_\ell(x)$

$$p(x) = \sum_{\ell=1}^{L} \pi_\ell \; p_\ell(x)$$

π_ℓ: mixture weight; $p_\ell(x)$: simple

$\pi_\ell \geq 0$, $\sum_{\ell=1}^{L} \pi_\ell = 1$

e.g. histogramm

$$p_\ell(x) = \mathbb{1}[x_{min} + (\ell-1)h \leq x < x_{min} + \ell \cdot h] \times$$
$$\times \text{uniform}(x_{min} + (\ell-1)h, x_{min} + \ell \cdot h)$$
$$= \mathbb{1}[\ldots] \frac{1}{h}$$

$$\pi_\ell = \frac{N_\ell}{N}$$

ex 2 Gaussian mixture model (GMM)

$$p_\ell(x) = \mathcal{N}(x \mid \mu_\ell, \sigma_\ell^2)$$

$$p(x) = \sum_{\ell=1}^{L} \pi_\ell \, \mathcal{N}(x \mid \mu_\ell, \sigma_\ell^2)$$

task: learn $\pi_\ell, \mu_\ell, \sigma_\ell^2$ for all $\ell = 1, \ldots, L$, s.t. $p(x) \approx p^*(x)$ according to TS

$\Rightarrow$ EM algorithm ("expectation-maximization")

alternating optimization
(a) learn μ_ℓ, σ_ℓ^2 with π_ℓ fixed
(b) -"- π_ℓ with μ_ℓ, σ_ℓ^2 fixed

alg: fix L as a hyperparameter

(0) initialization: select $\mu_\ell^{(0)}$ randomly (e.g. k-means++ strategy)

abbrev. $\sigma_\ell^2 =: \tau_\ell \quad \tau_\ell^{(0)} = 1 \quad \pi_\ell^{(0)} = \frac{1}{L}$

(1) for $t = 1, \dots, T$:

(a) $$\gamma_{i\ell} = p^{(t-1)}(x_i \in \text{cluster } \ell \mid x_i) = \frac{\pi_\ell^{(t-1)} \cdot p_\ell^{(t-1)}(x_i)}{\sum_{\ell'=1}^{L} \pi_{\ell'}^{(t-1)} \cdot p_{\ell'}^{(t-1)}(x_i)}$$

importance of $p_\ell(x)$ to explain $p(x_i)$

$$\gamma_{i\ell} = \frac{\pi_\ell^{(t-1)} \, N(x_i \mid \mu_\ell^{(t-1)}, \tau_\ell^{(t-1)})}{\sum_{\ell'=1}^{L} \pi_{\ell'}^{(t-1)} \, N(x_i \mid \mu_{\ell'}^{(t-1)}, \tau_{\ell'}^{(t-1)})}$$

(b) update μ_ℓ, τ_ℓ:

$$\mu_\ell^{(t)} = \frac{\sum_{i=1}^{N} \gamma_{i\ell} \, x_i}{\sum_{i=1}^{N} \gamma_{i\ell}} \qquad \tau_\ell^{(t)} = \frac{\sum_{i=1}^{N} \gamma_{i\ell} \, (x_i - \mu_\ell^{(t)})^2}{\sum_{i=1}^{N} \gamma_{i\ell}}$$

(c) $$\pi_\ell^{(t)} = \frac{1}{N} \sum_{i=1}^{N} \gamma_{i\ell}$$

learning $p(x)$ from TS in 1-D (contd.)

mixture models: superposition of simpler models

$$p(x) = \sum_{\ell=1}^{L} \pi_\ell \, p_\ell(x)$$

simple

mixture weights s.t. $\sum_\ell \pi_\ell = 1$, $\pi_\ell > 0$

(1) histogramm: $p_\ell(x)$ is uniform

- pick hyperparameter L
- define bins (e.g. regular grid) or learn bins (e.g. recursive subdivision, $\Rightarrow$ density tree)

$$\Rightarrow p_\ell(x) = \text{uniform}(\text{bin}_\ell) = \frac{\mathbb{1}[x \in \text{bin}_\ell]}{\text{volume}(\text{bin}_\ell)}$$

(in 1D volume$(\text{bin}_\ell) = h$ bin width)

- learn $\pi_\ell = \frac{N_\ell}{N}$ ← # instances in bin$_\ell$

(2) Gaussian mixture model, $p_\ell(x) = \mathcal{N}(\mu_\ell, \Sigma_\ell)$

- pick L
- train $\pi_\ell, \mu_\ell, \Sigma_\ell$ by the EM algorithm (last lecture)

(3) kernel density estimation $p_\ell(x) = \mathcal{N}(\mu_\ell, \sigma^2 \mathbb{I})$ or any other simple distribution

- $L = N$: one component per training instance
- $\mu_\ell = x_\ell$, $\pi_\ell = \frac{1}{N}$, σ^2 chosen as hyperparameter