Machine Learning (Spring 2020) (Due: 16.04.2020 23:55)

Assignment 2

Instructor: Dr. Sergey Kosov *TA:* Yuhou Zhou

Course Policy: Read all the instructions below carefully before you start working on the assignment, and before you make a submission.

- The homework assignments are for practice purpose. The grade from your homework will not affect your final grade of the course.
- Please submit your answer sheet, either by a scanned copy or a typeset PDF file, to Moodle before the deadline.
- No late submission is accepted.
- You can do this assignment in groups of 2. Please submit no more than one submission per group.

Problem 1: K-nearest Neighbors (2.5+2.5+5=10 points)

(a) Given a 2 dimensional data set:

$$T = \{(2,3)^T, (5,4)^T, (9,6)^T, (4,7)^T, (8,1)^T, (7,2)^T\}$$

Construct a balanced kd-tree.

Order nodes by their x values ascendently.

$(2,3)^T, (4,7)^T, (5,4)^T, (7,2)^T, (8,1)^T, (9,6)^T.$

Find the median of x. Here we get either $(5,4)^T$ or $(7,2)^T$.

I take $(7,2)^T$ as the case.

$(2,3)^T, (4,7)^T, (5,4)^T, (7,2)^T, (8,1)^T, (9,6)^T$

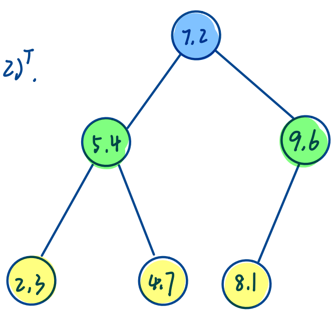
We then order 2nd layer nodes by y.

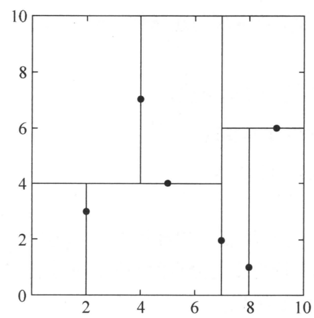
$(2,3)^T, (5,4)^T, (4,7)^T, (7,2)^T, (8,1)^T, (9,6)^T$

or

$(2,3)^T, (5,4)^T, (4,7)^T, (7,2)^T, (8,1)^T, (9,6)^T$

Therefore a balanced k-d tree:

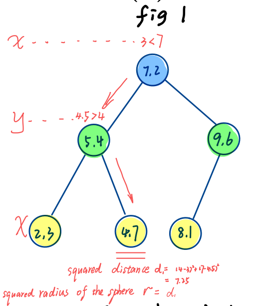




partitioned space

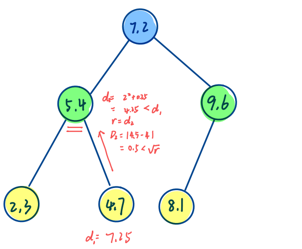
(b) Use the kd-tree constructed in problem (a) to find the nearest point of $x = (3, 4.5)^T$.

fig 1



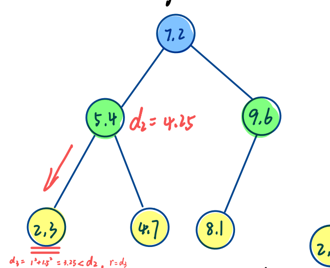
First, go from the root to leaves. It goes left or right based on the point is less or greater than the current node in the split dimension.

fig 2



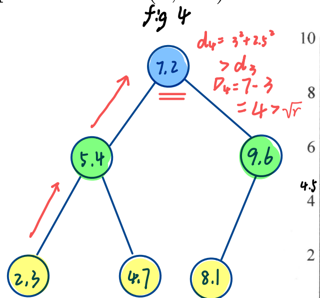
After reaching leaves, start to go back. Because $D_2 > d_2$, the sphere reach the other side of the hyperplane. We go to the other side of the current node. See fig 3.

fig 3

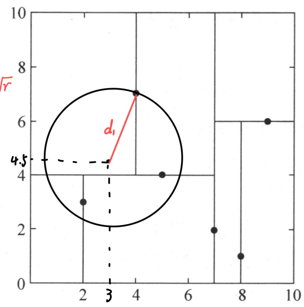


Currently d_3 has the nearest distance. Since we reach a leaf node, we go up again. See fig 4.

fig 4



We reach the root node, $d_4 > d_3$ and $D_4 > \sqrt{r}$ so we don't need to go to the other side. Search finished. The nearest node is $(2,3)$.



phase 1

phase 2

https://en.m.wikipedia.org/wiki/K-d_tree#Nearest_neighbour_search

(c) Show that the k-nearest-neighbour density model defines an improper distribution whose integral over all space is divergent.

If we want to classify a new point x, we draw a sphere centered on X containing precisely K points irrespective their class.

The unconditional density is given by

$$p(x) = \frac{k}{N} \cdot \frac{1}{V}$$

where K is the number of points in the sphere. N is the number of total observations. V is the volume of the sphere.

In the n dimension, the volume of the n-ball:

$$V = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} R^n$$

where Γ is Leohard Euler's gamma function. It extends the factorial function to non-integer argument. R is the radius of the n-ball.

We order the points according to their distances to the data point: $x < X_1 < X_2 < X_3 < \cdots < X_k < \cdots < X_N$

$R = \| X_k - x \|$

Substitute V, R in $P(x)$:

$$P(x) = \frac{K}{N} \frac{\Gamma(2n+1)}{\pi^{n/2}} \frac{1}{\| X_k - X \|^n}$$

The integration over domain $\mathbb{R}^n$.

$$\int_{\mathbb{R}^n} p(x) dx = \frac{K}{N} \frac{\Gamma(2n+1)}{\pi^{n/2}} \int_{\mathbb{R}^n} \frac{1}{\| x_k - x \|^n} dx$$

If we assume data points are in 1 dimension, i.e., $n=1$.

then the above:

$$\frac{K}{N} \frac{\Gamma(3)}{\pi^{1/2}} = \frac{K}{N} \frac{\frac{1}{2} \cdot \pi^{1/2}}{\pi^{1/2}} = \frac{K}{2N}$$

$$\int_{-\infty}^{+\infty} \frac{1}{\| X_k - X \|} dx = \int_{-\infty}^{x_1} \frac{1}{\| X_k - X \|} dx + \int_{x_1}^{+\infty} \frac{1}{\| X_k - X \|} dx$$

because $x < x_1 < x_k$,

$$= \int_{-\infty}^{x_1} \frac{1}{(x_k - x)} dx + \int_{x_1}^{+\infty} \frac{1}{\| x_k - x \|} dx$$

$$= \Big[\ln (X_k - x) \Big]_{-\infty}^{x_1} + \int_{x_1}^{+\infty} \frac{1}{\| x_k - x \|} dx$$

because $\frac{1}{\| x_k - x \|} \geqslant 0$. $\int_{x_1}^{+\infty} \frac{1}{\| x_k - x \|} dx \geqslant 0$.

So $\int_{-\infty}^{+\infty} p(x) dx = +\infty$.

This remains true for $x \in \mathbb{R}^n$, but needs a bit more computing skills.

Problem 2: Gaussian Mixture Model

(5+5+5+5=20 points)

(a) Consider a Gaussian mixture model in which the marginal distribution $p(\mathbf{z})$ for the latent vatiable is given by $p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$, and the conditional distribution $p(\mathbf{x}|\mathbf{z})$ for the observaed variable is given by $p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$. Show that the marginal distribution $p(\mathbf{x})$, obtained by summing $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ over all possible values of $\mathbf{z}$, is a Gaussian mixture of the form $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Note: $\mathbf{z}$ uses 1-of-K representation.

$$p(x) = \sum_{z} p(z) p(x|z)$$

$$= \sum_{z} \prod_{k=1}^{K} \left[\pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \right]^{z_k}$$

Because z uses 1-of-K representation.

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

$z = [z_1, z_2, \dots, z_K]$, where only one element of z is 1, and 0 otherwise.

* For instance: If $z = [0, 1, 0]$,

$$p(z) = \prod_{k=1}^{3} \pi_k^{z_k} = \pi_1^0 \pi_2^1 \pi_3^0 = 1 \cdot \pi_2 \cdot 1 = \pi_2$$

* If $K = 3$, $\sum_z$ means for each loop, you use one of $z = [1, 0, 0]$, $z = [0, 1, 0]$, $z = [0, 0, 1]$.

latent data

$$\begin{array}{c|ccccc} & 1 & \cdots & k & \cdots & K \\ \hline z_1 & & & \vdots & & \\ \vdots & & & & & \\ z_n & \cdots & & z_{nk} & & \\ \vdots & & & & & \\ z_N & & & & & \end{array}$$

(b) Verify that maximization of the complete-data log likelihood

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

for a Gaussian mixture model leads to the result that the means and covariances of each component are fitted independently to the corresponding group of data points, and the mixing coefficients are given by the fractions of points in each group.

$\max_{\mu_k} \ln p$:

$$\frac{\partial \ln p}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \sum_{n=1}^{N}\sum_{j=1}^{K} z_{nj}\{\ln \pi_j + \ln N(x_n|\mu_j, \Sigma_j)\}$$

$$= \sum_{n=1}^{N} \frac{\partial}{\partial \mu_k} z_{nk} \ln N(x_n|\mu_k, \Sigma_k)$$

$$= \sum_{x_n \in X_k} \frac{\partial}{\partial \mu_k} \ln N(x_n|\mu_k, \Sigma_k)$$

Where X_k contains x_n assigned to the group k.

It's the same for $\max_{\Sigma_k} \ln p$

$\max_{\pi_k} \ln p$

s.t. $\sum_{k=1}^{K} \pi_k - 1 = 0$:

$$\mathcal{L}(\pi_k, \lambda) = \ln p + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \frac{\partial}{\partial \pi_k}\left[\sum_{n=1}^{N} z_{nk} \ln \pi_k + \lambda(\pi_k - 1)\right]$$

$$= \sum_{n=1}^{N} \frac{z_{nk}}{\pi_k} + \lambda$$

$$\begin{cases} \sum_{n=1}^{N} \frac{z_{nk}}{\pi_k} + \lambda = 0 \\ \sum_{k=1}^{K} \pi_k - 1 = 0 \end{cases} \Rightarrow \begin{cases} \pi_k \lambda = -\sum_{n=1}^{N} z_{nk} \\ \sum_{k=1}^{K} \pi_k = 1 \end{cases}$$

$$\Rightarrow \begin{cases} \sum_{k=1}^{K} \pi_k \lambda = -\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk} \\ \sum_{k=1}^{K} \pi_k = 1 \end{cases} \Rightarrow \begin{cases} \lambda = -N \\ \pi_k = \frac{1}{N}\sum_{n=1}^{N} z_{nk} = \frac{N_k}{N} \end{cases}$$

(c) Show that if we maximize

$$\mathbb{E}_{\mathbf{z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma(z_{nk})\{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

with respect to $\boldsymbol{\mu}_k$ while keeping the responsibilities $\gamma(z_n k)$ fixed, we obtain the closed form solution given by

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n$$

$$\frac{\partial}{\partial \mu_k} \mathbb{E}_z[\ln p] = \sum_{n=1}^{N}\sum_{j=1}^{K} \gamma(z_{nj}) \frac{\partial}{\partial \mu_k} \ln N(x_n|\mu_j, \Sigma_j) \quad \cdots ①$$

$$\because N(x|\mu, \Sigma) = \frac{1}{2\pi^{D/2}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

$\therefore$ ① becomes

$$\sum_{n=1}^{N} \gamma(z_{nj}) \frac{\partial}{\partial \mu_k} \sum_{j=1}^{K} -\frac{1}{2}(x_n - \mu_j)^T \Sigma^{-1}(x_n - \mu_j)$$

because only μ_k is variable of μ_j, where $j = 1 \ldots k \ldots K$.

so only the terms with μ_k is not 0, when $\frac{\partial}{\partial \mu_k}$.

$$-\sum_{n=1}^{N} \gamma(z_{nk}) \Sigma^{-1}(x_n - \mu_k) \quad \cdots ②$$

Set ② $= 0$:

$$\sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k) = 0$$

$$\sum_{n=1}^{N} \gamma(z_{nk})\mu_k = \sum_{n=1}^{N} \gamma(z_{nk}) x_n$$

$$N_k \mu_k = \sum_{n=1}^{N} \gamma(z_{nk}) x_n$$

$$\mu_k = \frac{1}{N_k}\sum_{n=1}^{N} \gamma(z_{nk}) x_n$$

where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$.

(d) Consider a density model given by a mixture distribution

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|k)$$

and suppose that we partition the vector $\mathbf{x}$ into two parts so that $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$. Show that the conditional density $p(\mathbf{x}_b|\mathbf{x}_a)$ is itself a mixture distribution and find expressions for the mixing coefficients and for the component densities.

The question asks for three parts.

1> $p(x_b|x_a)$ has the same form of $p(x)$.

2> Expression of $\pi_{k(x_b|x_a)}$.

3> Component densities of $p(x_b|x_a)$.

$$p(x_b|x_a) = \frac{p(x_b, x_a)}{p(x_a)} = \frac{p(x)}{p(x_a)} = \sum_{k=1}^{K} \frac{\pi_k}{p(x_a)} p(x|k) \quad \cdots \text{①}$$

$$p(x_a) = \sum_{j=1}^{K} \pi_j p(x_a|j) \quad \cdots \text{②}$$

Substitute ② to ①:

$$p(x_b|x_a) = \sum_{k=1}^{K} \underbrace{\frac{\pi_k}{\sum_{j=1}^{K} \pi_j p(x_a|j)}}_{\pi_{k(x_b|x_a)}} \underbrace{p(x|k)}_{\text{component density}}$$