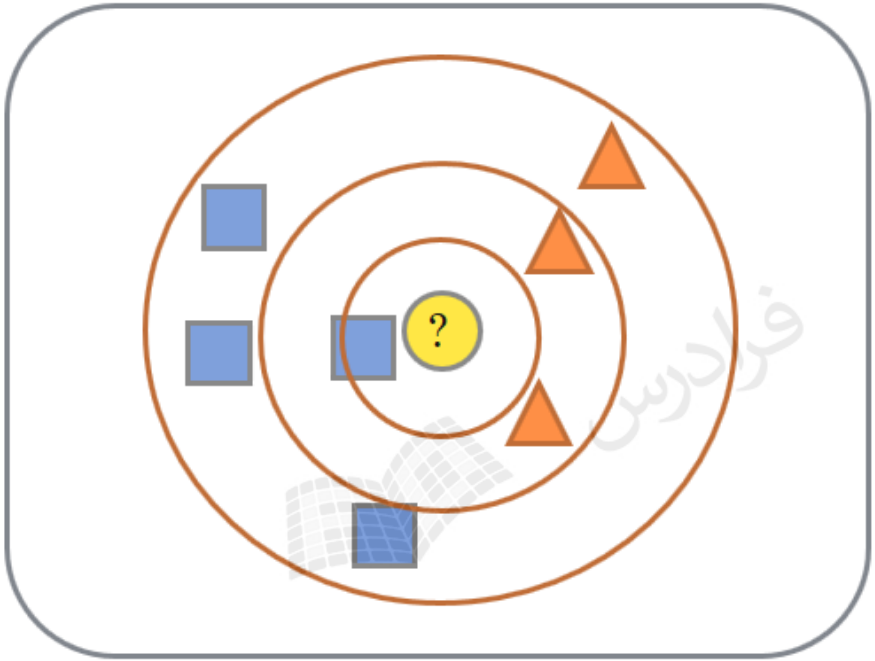
فرادرس
فراتر از یک کلاس درس
www.faradars.org

آموزش یادگیری ماشین (Machine Learning)
(تئوری – عملی) - بخش دوم

درس دوم: دسته‌بندی K نزدیک‌ترین همسایه

مدرس:
فرشید شیرافکن
دانشجوی دکترای بیو انفورماتیک
دانشگاه تهران

KNN



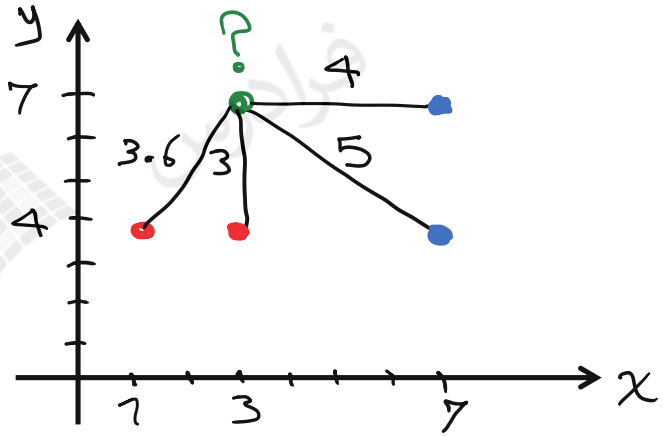
- $k = 1$: square class
- $k = 3$: triangle class
- $k = 7$: square class

مثال

x	y	class
7	7	False
7	4	False
3	4	True
1	4	True
3	7	? True

x=3 ، y=7

K=3



$$\sqrt{(3-7)^2+(7-4)^2}=5$$

$$\sqrt{(3-3)^2+(7-4)^2}=3 \quad \text{True}$$

$$\sqrt{(3-7)^2+(7-4)^2}=3.6 \quad \text{True}$$

$$\sqrt{(3-7)^2+(7-7)^2}=4 \quad \text{False}$$

مثال

ویژگی

Customer	X	Y	Z	C
ali	35	35	3	No ✓
sara	22	50	2	Yes ✓
farid	63	200	1	No
taha	59	170	1	No
omid	25	40	4	Yes ✓

reza	**37**	**50**	**2**	**?**

K=3

$\sqrt{(35-37)^2+(35-50)^2+(3-2)^2} = 75.16$

$\sqrt{(22-37)^2+(50-50)^2+(2-2)^2} = 75$

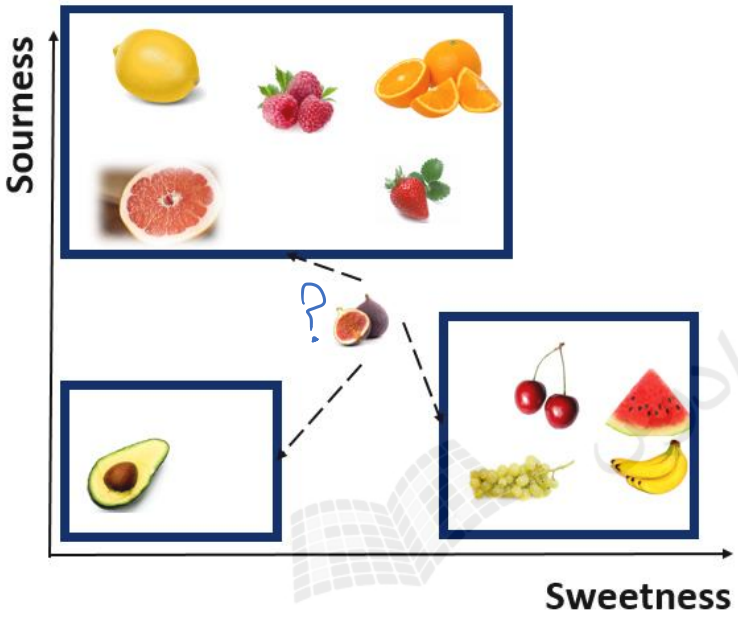
152.23

122

75.74

→ yes

مثال



3NN

Fruit	Sweetness	Sourness	Fruit Type
Lemon	1	9	Sour
Grapefruit	2	8	Sour
Orange	3	7	Sour
Raspberry	2	8	Sour
Cherry	6	4	Sweet
Banana	9	1	Sweet
Grapes	8	2	Sweet
Watermelon	9	1	Sweet
Avocado	1	1	None
Strawberry	5	5	Sour

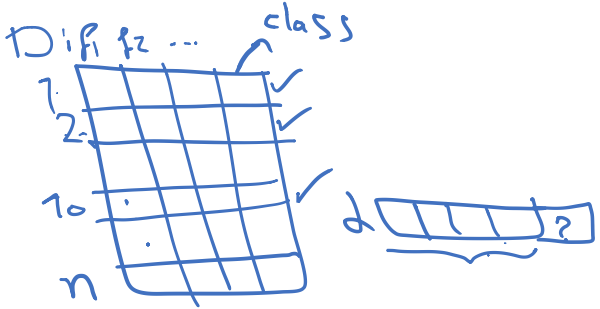
Fruit	Sweetness	Sourness
Fig	7	3

?

الگوریتم KNN

KNN(D , d , k)

1. Compute the distance between d and every example in D;
2. Choose the k examples in D that are nearest to d, denote the set by P ($\subseteq$ D);
3. Assign d the class that is the most frequent class in P (or the majority class).



تکنیک‌های instance-based

Model-based learning techniques

Use the input data

$$\begin{bmatrix} x_{1,0} & x_{1,1} & \dots & x_{1,n} \\ x_{2,0} & x_{2,1} & \dots & x_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m,0} & x_{m,1} & \dots & x_{m,n} \end{bmatrix} \text{ and } \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{bmatrix}$$

To learn a set of parameters

$$\begin{bmatrix} \theta_0 & \theta_1 & \dots & \theta_n \end{bmatrix}$$

Which yield a **generalized** function

$$f(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

Capable of predicting values or classes on new input data

$$f(x_i) = 39$$

$$f(x_j) = 1$$

Instance-based learning techniques

Store the input data

$$\begin{bmatrix} x_{1,0} & x_{1,1} & \dots & x_{1,n} \\ x_{2,0} & x_{2,1} & \dots & x_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m,0} & x_{m,1} & \dots & x_{m,n} \end{bmatrix} \text{ and } \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{bmatrix}$$

When asked to predict a new value (a query)

$$y_i = ?$$

Search for similar data points previously stored

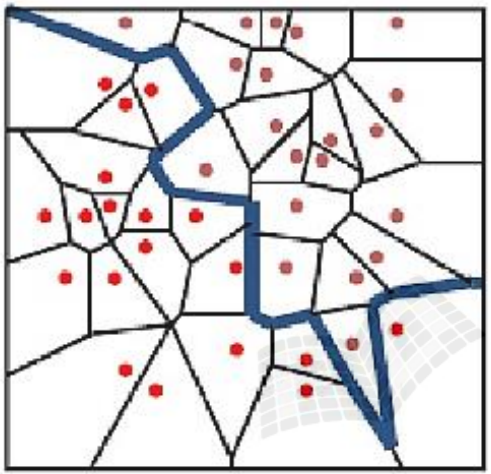
$$\begin{bmatrix} x_{4,1} & x_{4,2} & \dots & x_{4,n} \\ x_{9,1} & x_{9,1} & \dots & x_{9,n} \\ x_{15,1} & x_{15,1} & \dots & x_{15,n} \end{bmatrix} \text{ and } \begin{bmatrix} y_4 \\ y_9 \\ y_{15} \end{bmatrix}$$

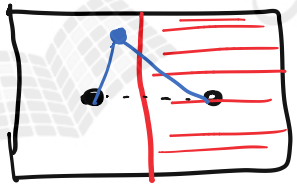
And use them to generate your prediction

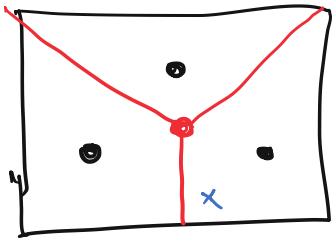
$$y_i = \frac{y_4 + y_9 + y_{15}}{3}$$

مرز تصمیم – نمودار وُرُنوی

decision boundaries – Voronoi diagram



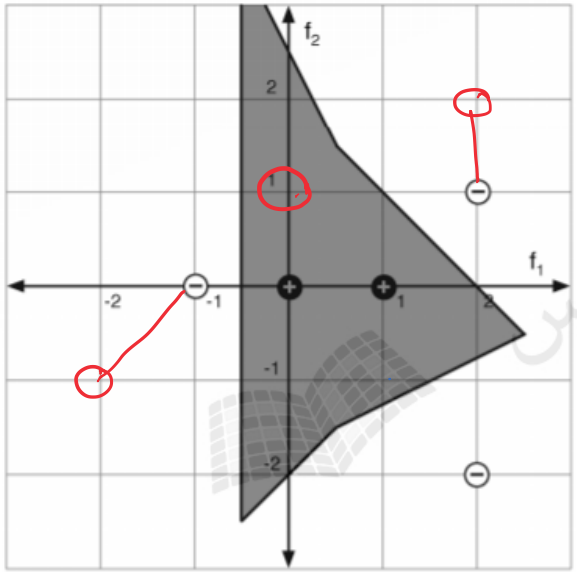


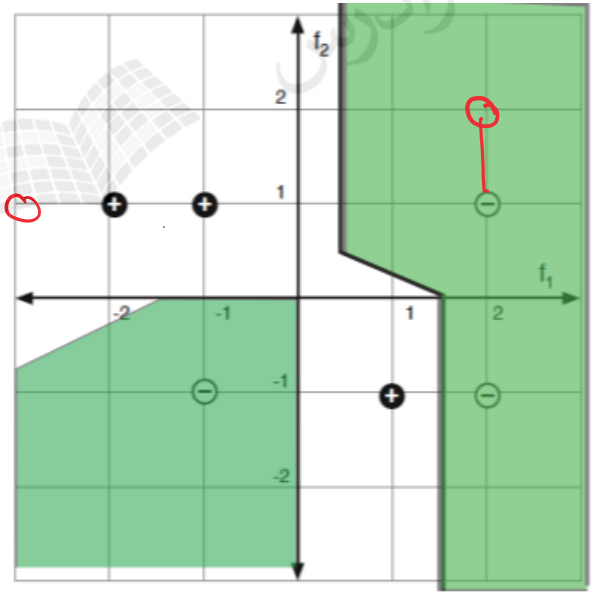


فرادرس

مثال

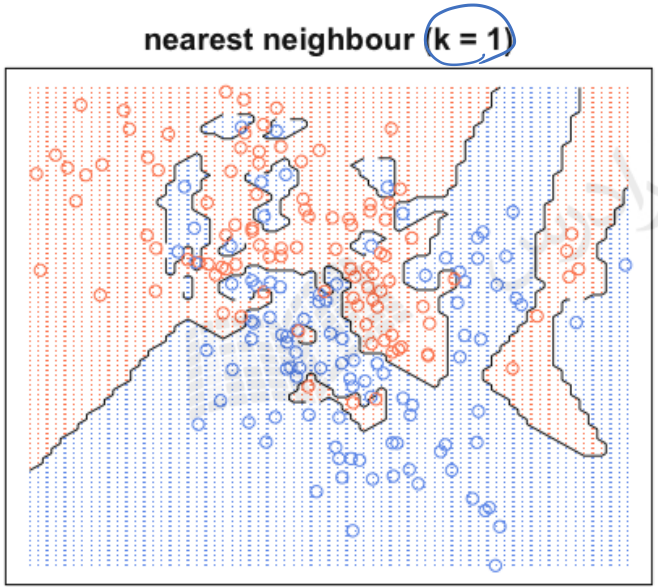
مرزهای تصمیم برای 1-NN

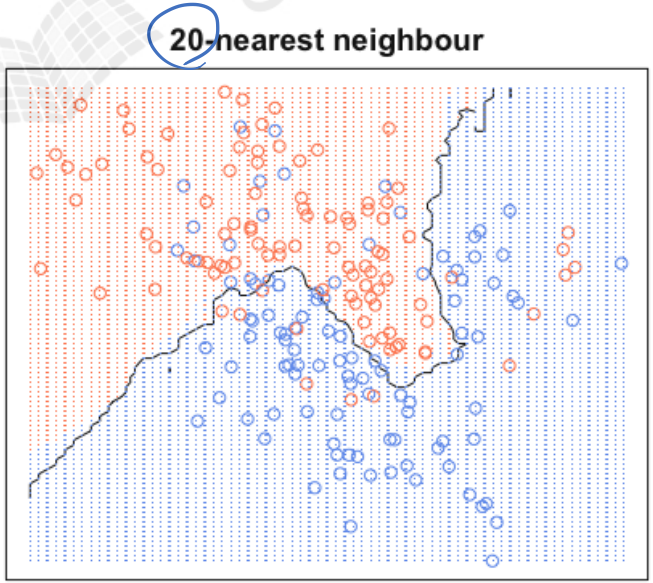




تاثیر اندازه k در مرز تصمیم

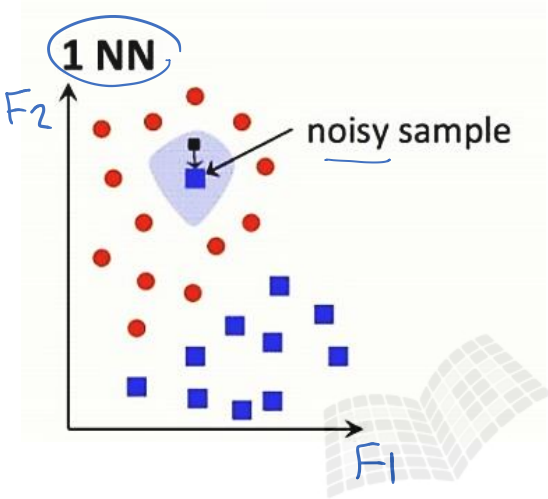
مقدار k کوچک باعث پیچیده شدن مرز تصمیم می‌شود. با افزایش مقدار k، مرزهای کلاس‌ها روان‌تر می‌شود.





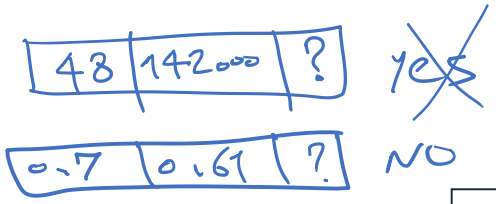
اندازه k

اگر k را خیلی کوچک در نظر بگیریم، به نقاط نویزی حساس می‌شود.



عموما مقدار k را برای دسته‌بندی‌های دودویی (binary classification)، فرد در نظر می‌گیرند.

استانداردسازی



Min–max normalization: $$X_s = \frac{X - Min}{Max - Min}$$

$\frac{25-20}{60-20} = \frac{5}{40}$ $\frac{22000-18000}{220000-18000}$

Age	Loan	Default	Distance
25	$40,000	N	102000
35	$60,000	N	82000
45	$80,000	N	62000
20	$20,000	N	122000
35	$120,000	N	22000
52	$18,000	N	124000
23	$95,000	Y	47000
40	$62,000	Y	80000
60	$100,000	Y	42000
48	$220,000	Y	78000
33	$150,000	Y	8000

20 – 60 18000 – 220000

Age	Loan	Default	Distance
0.125	0.11	N	0.7652
0.375	0.21	N	0.5200
0.625	0.31	N	0.3160
0	0.01	N	0.9245
0.375	0.50	N	0.3428
0.8	0.00	N	0.6220
0.075	0.38	Y	0.6669
0.5	0.22	Y	0.4437
1	0.41	Y	0.3650
0.7	1.00	Y	0.3861
0.325	0.65	Y	0.3771

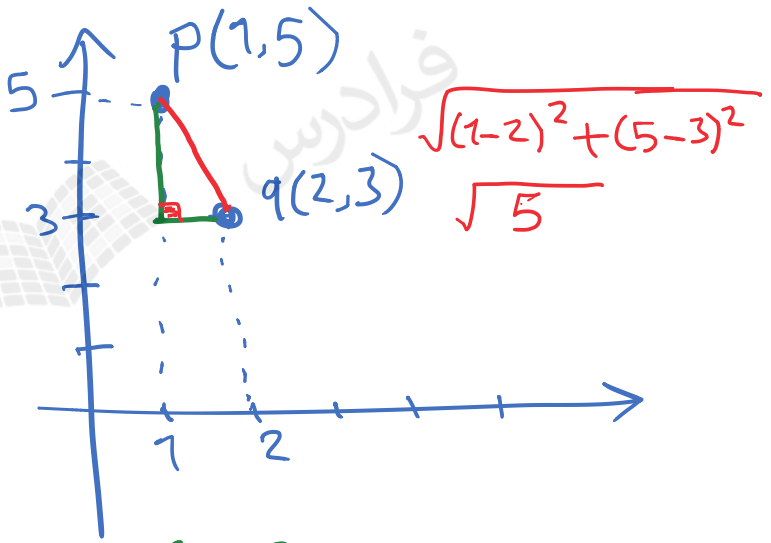
معیارهای شباهت

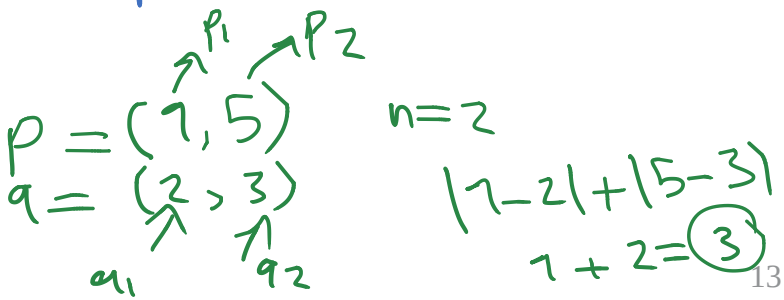
Minkowski: $$d(p,q) = \left(\sum_{i=1}^{n} |p_i - q_i|^x\right)^{\frac{1}{x}}$$

x=2: Euclidian $$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

x=1: Manhatan $$d(p,q) = \sum_{i=1}^{n} |p_i - q_i|$$

city_block



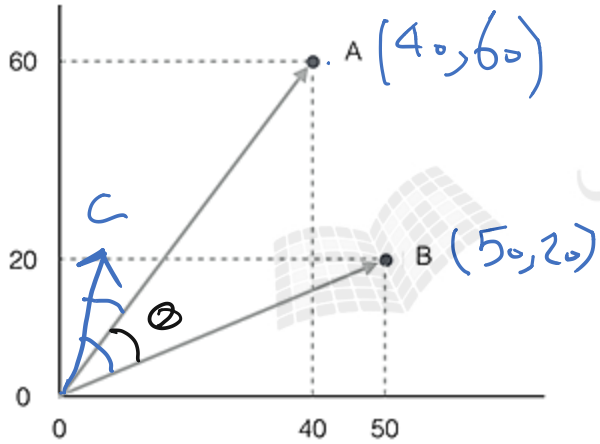


معیار شباهت کسینوس

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

$$A \cdot B = |A|\,|B| \cos\theta$$

$$\cos\theta = \frac{40 \times 50 + 60 \times 20}{\sqrt{40^2 + 60^2} \times \sqrt{50^2 + 20^2}} = \bigcirc$$



مثال

q: **gold silver truck**

- d1: Shipment of gold damaged in a fire
- d2: Delivery of silver arrived in a silver truck
- d3: Shipment of gold arrived in a truck

$$Sim(q, d_1) = \frac{1 \times 1}{\sqrt{3} \times \sqrt{7}}$$

$$Sim(q, d_2) = \frac{1 \times 2 + 1 \times 1}{\sqrt{3} \times \sqrt{10}}$$

$$Sim(q, d_3) = \frac{1 \times 1 + 1 \times 1}{\sqrt{3} \times \sqrt{7}}$$

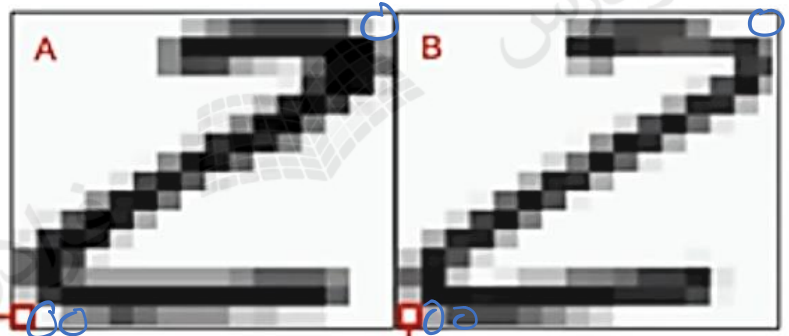
Terms	d1	d2	d3	q
a	1	1	1	0
arrived	0	1	1	0
damaged	1	0	0	0
delivery	0	1	0	0
fire	1	0	0	0
gold	1	0	1	1
in	1	1	1	0
of	1	1	1	0
shipment	1	0	1	0
silver	0	2	0	1
truck	0	1	1	1

A = , q =

تشخیص ارقام دست‌نویس

MNIST

0 — 9

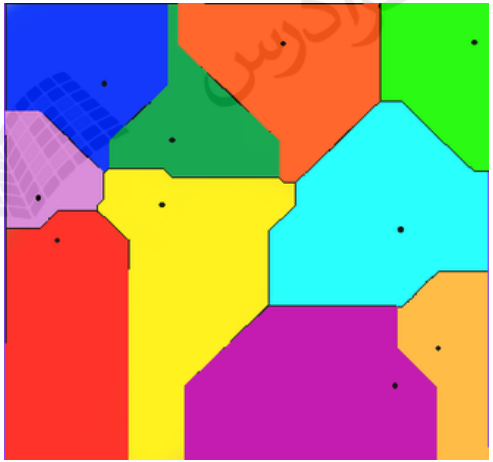


$$D(A,B)=\sqrt{\sum_r \sum_c (A_{r,c}-B_{r,c})^2}$$

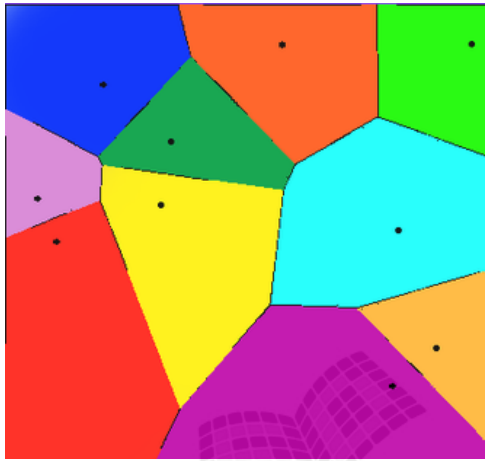
KNN
K=7

2 2 2 3 7 1 8

تاثیر معیار شباهت در مرز تصمیم



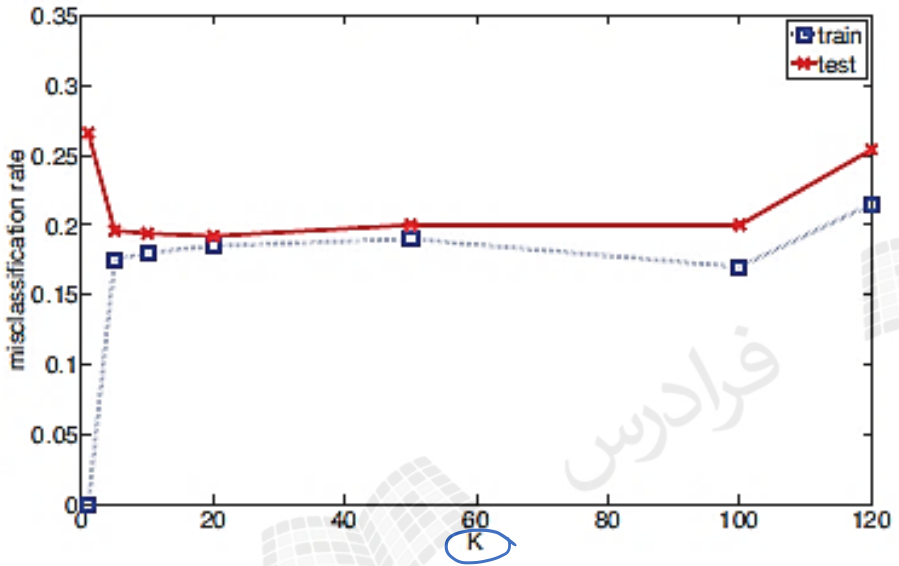
منهتن



اقلیدسی

فرادرس

استراتژی انتخاب k



KNN K=?

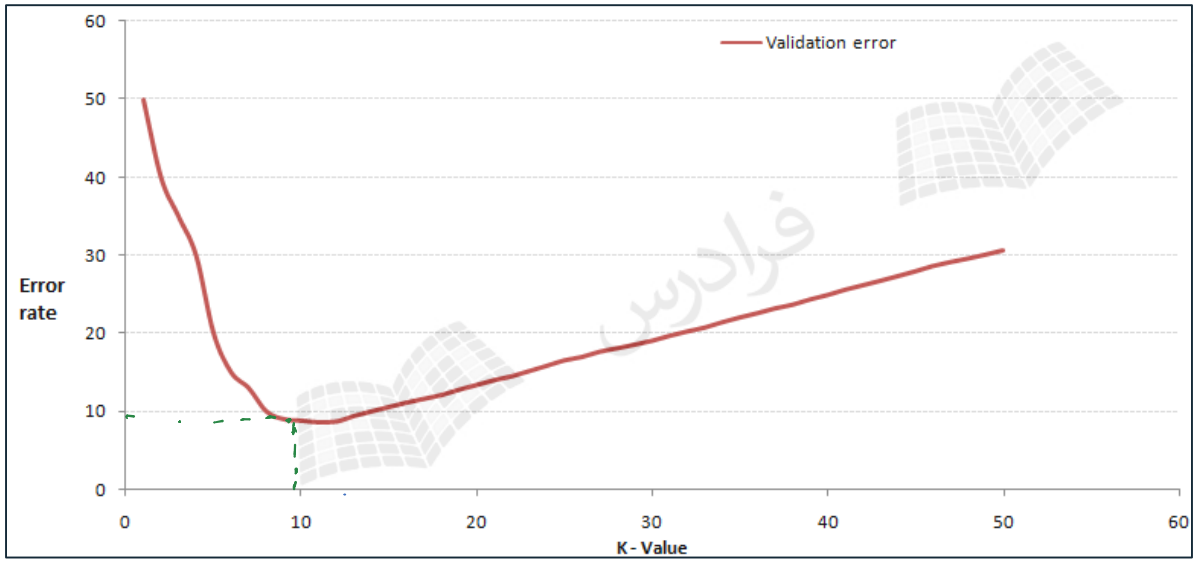
انتخاب k با استفاده از cross validation.

where *K* is small, the model is complex and hence we overfit.

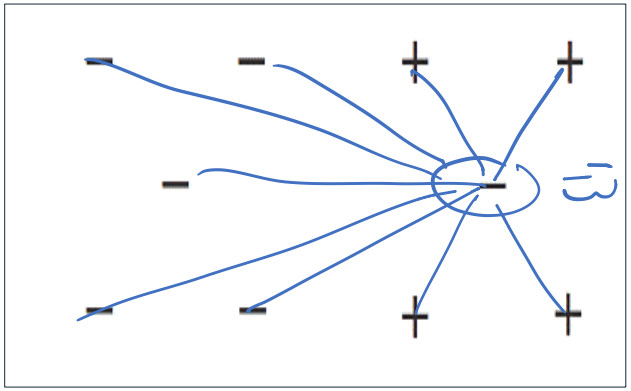
where *K* is large, the model is simple and we underfit.

مثال

Validation error is the least when the value of k is 10.



مثال



خطای LOOCV؟

LOOCV: leave one out cross validation

K=1

1NN

5/10

3NN 1/10

مزایای KNN

- پیاده‌سازی ساده
- صفر بودن هزینه مرحله آموزش
- عدم نیاز به پیش‌پردازش داده (معمولا)
- تفسیر ساده

فرادرس

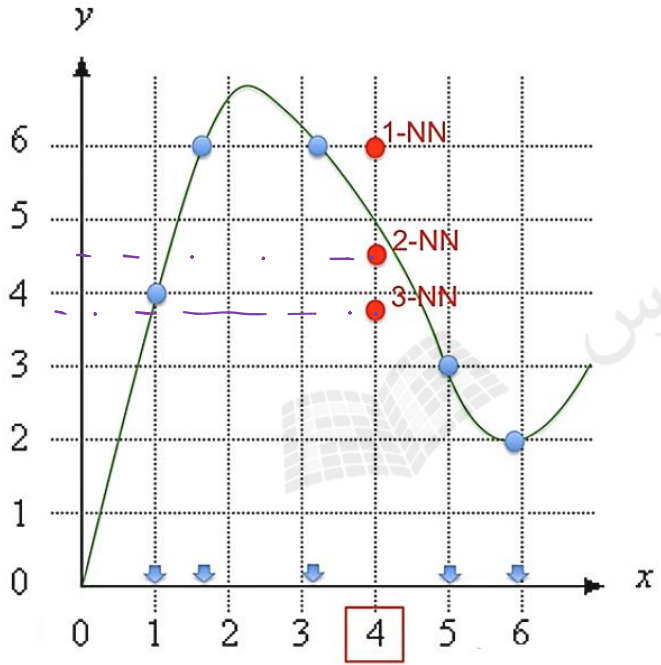
معایب KNN

- هزینه‌بر بودن تست داده جدید
- کاهش دقت به‌علت داده‌های پرت و نویز
- داشتن بایاس برای ابعاد بزرگ

فرادرس

استفاده از KNN برای پیشگویی عددی

KNN Regression



x	y
1	4
1.8	6
→ 3.1	6
→ 5	3
→ 6	2
4	?

y = f(x)

1NN : 6

2NN : $\frac{6+3}{2} = 4.5$

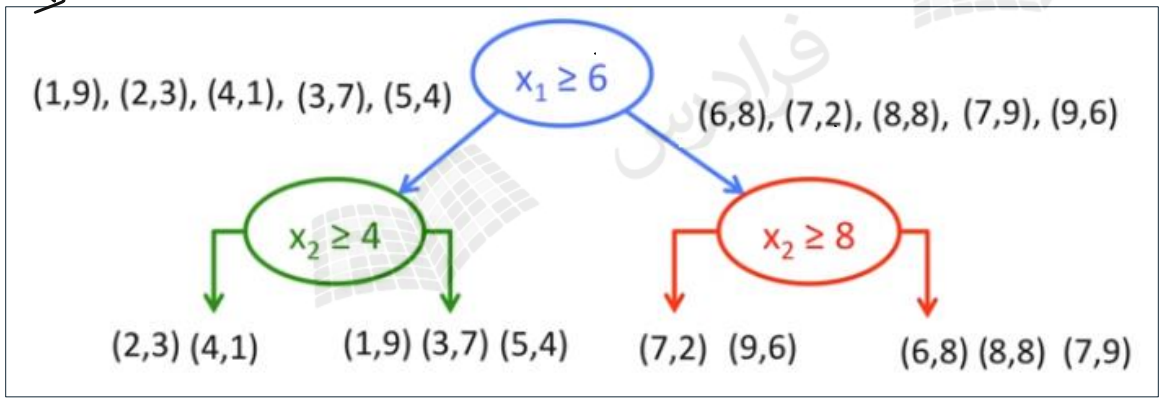
3NN : $\frac{6+3+2}{3} = \frac{11}{3}$

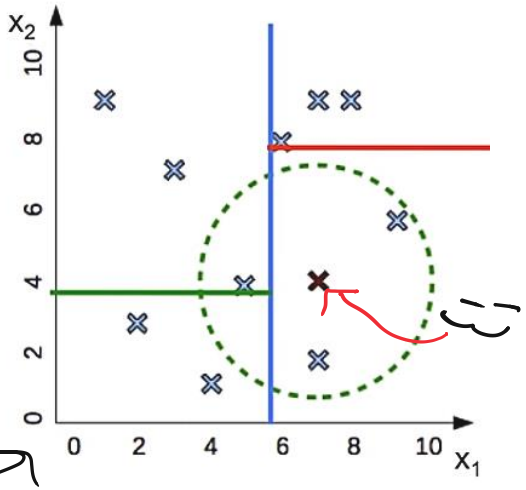
استفاده از kd-tree

می‌توان از kd-tree برای سریع شدن جستجوی نزدیک‌ترین همسایه در KNN استفاده کرد.

x_1 x_2

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





این اسلایدها بر مبنای نکات مطرح شده در فرادرس
«آموزش یادگیری ماشین (Machine Learning) (تئوری - عملی) – بخش دوم»
تهیه شده است.

برای کسب اطلاعات بیشتر در مورد این آموزش به لینک زیر مراجعه نمایید.

faradars.org/fvdm94062