Classification

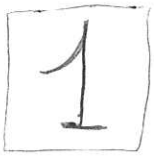
Use an object's attributes (features) to identify which category (or class) it belongs to.

Famous applications

- spam detection (spam / ham)
- image classification (cat / dog, plane / bird, ...)

{ binary classification

- handwritten digit recognition: 0/1/2/3.../9



- who wrote the "disputed federalist papers"
  Madison / Hamilton / John Jay

{ multiclass classification

Machine Learning Classification Algorithms:

- knn
- regression
- decision trees / random Forests
- Naive Bayes
- ⋮

{ implemented in the scikit-learn library

k-nearest neighbors (knn)

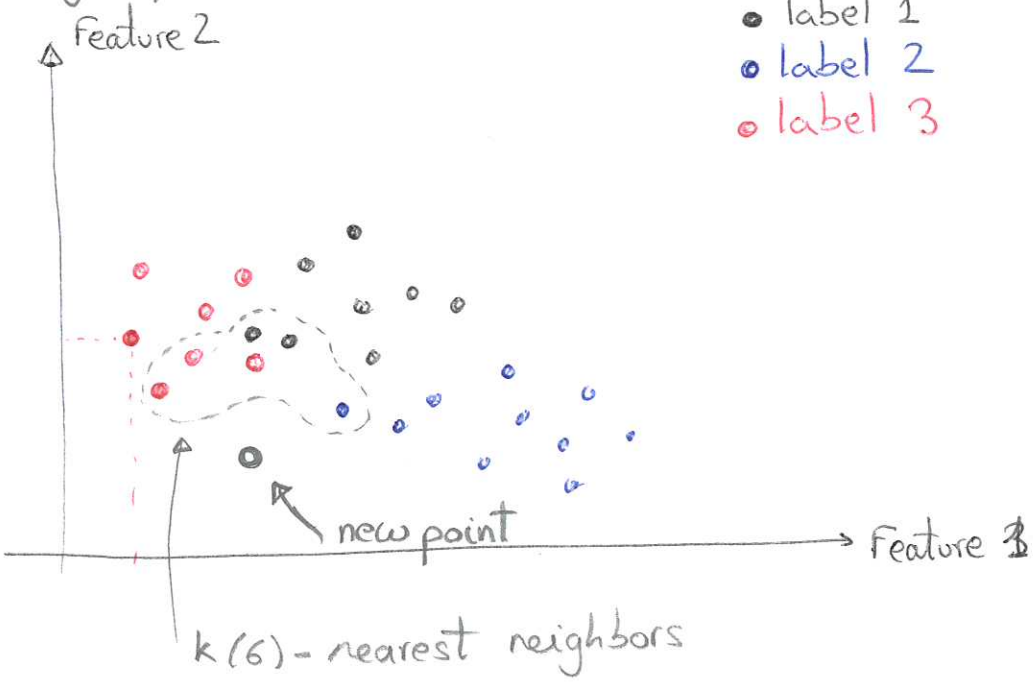
hyper-parameters: k (number of neighbors); a parameter that is set before the fitting process begins.

(i) fit: load data (aka train)

(ii) predict: { uniform / distance

uniform (majority rule)

(k=6)



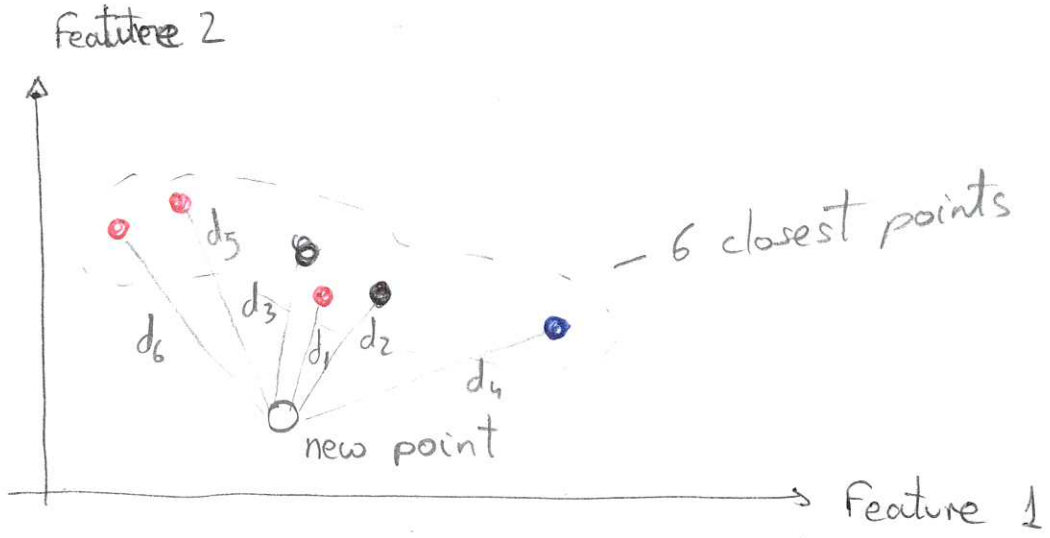
Probabilities:

$$P_1 = \frac{2}{6} = \frac{1}{3}$$

$$P_2 = \frac{1}{6}$$

$$P_3 = \frac{3}{6} = \frac{1}{2} \Longrightarrow \boxed{\text{prediction: } 3}$$

distance (nearer neighbors contribute more to the prediction)



distances

$d_1 = 1$

$d_2 = 1.25$

$d_3 = 1.5$

$d_4 = 2$

$d_5 = 5$

$d_6 = 6$

NOTE:

we divide by $\frac{1}{d_1} + \frac{1}{d_2} + \dots + \frac{1}{d_6}$

because probabilities have to sum up to 1

$P_1 + P_2 + P_3 = 1$

probabilities

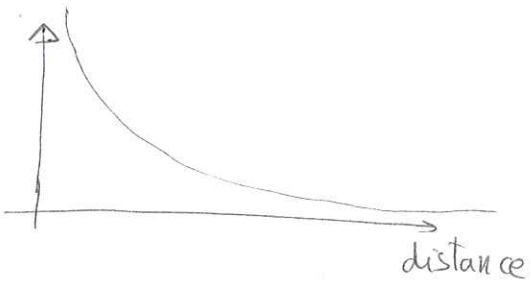
$$P_1 = \frac{\frac{1}{d_2} + \frac{1}{d_3}}{1/d_1 + 1/d_2 + \dots + 1/d_6} = \frac{1.47}{3.33} \Rightarrow \text{prediction: } 1$$

$$P_2 = \frac{1/d_4}{1/d_1 + 1/d_2 + \dots + 1/d_6} = \frac{0.5}{3.33}$$

$$P_3 = \frac{1/d_1 + 1/d_5 + 1/d_6}{1/d_1 + 1/d_2 + \dots + 1/d_6} = \frac{1.37}{3.33}$$

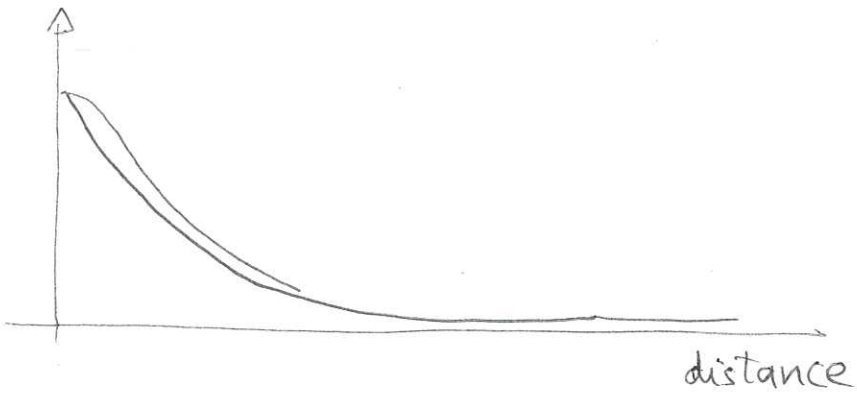
By default, distance-knn uses

$$\frac{1}{\text{distance}}$$

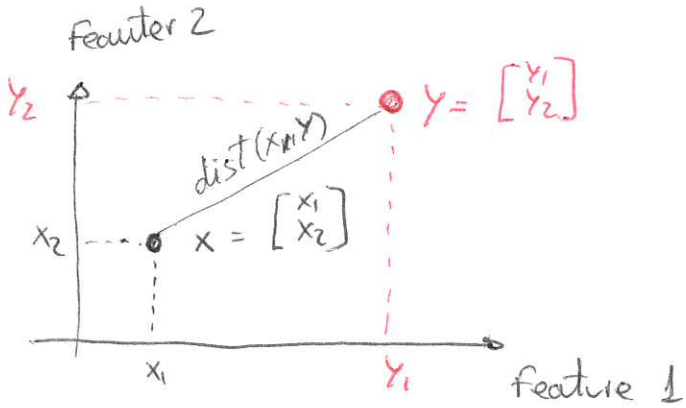


Other functions can be used. For example,

$e^{-\text{distance}}$.



What distance does knn use?



- Euclidean distance: $dist(x,y) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2}$
- Taxicab distance: $dist(x,y) = |x_1-y_1| + |x_2-y_2|$
- p-distance (minkowski): $dist(x,y) = \sqrt[p]{(x_1-y_1)^p+(x_2-y_2)^p}$
- cosine distance: $dist(x,y) = \dfrac{x_1y_1 + x_2y_2}{\sqrt{x_1^2+x_2^2}\ \sqrt{y_1^2+y_2^2}}$

⋮