Instance-Based Learning

The k-NN algorithm: simple application

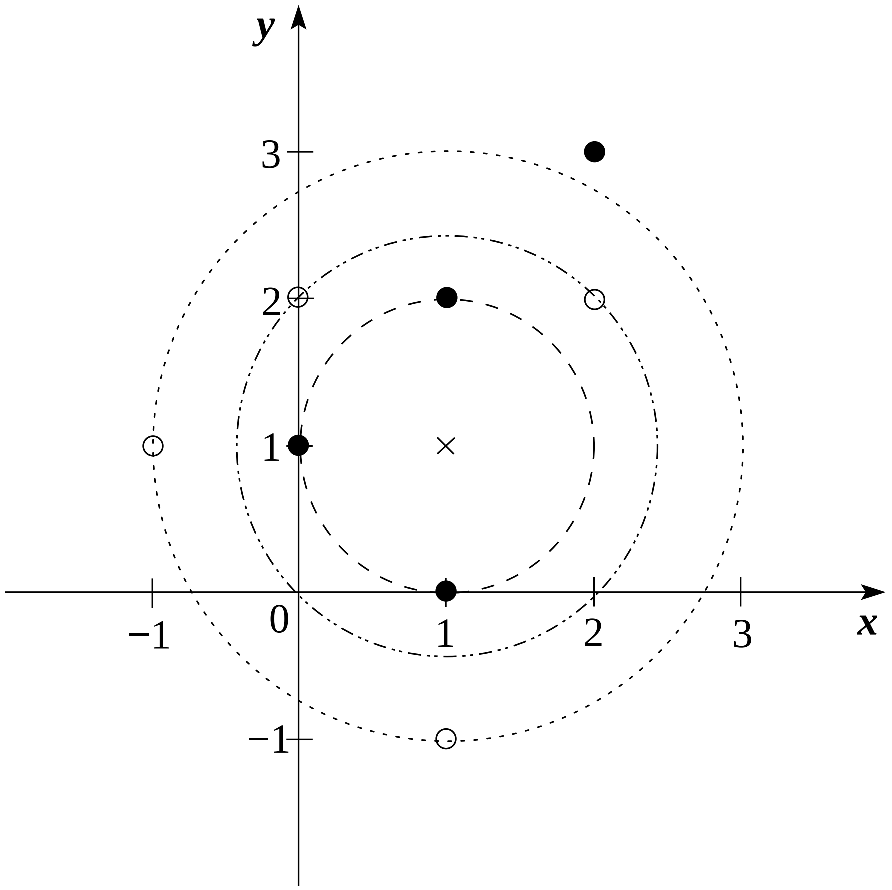
CMU, 2006 fall, final exam, pr. 2

Consider the training set in the 2-dimensional Euclidean space shown in the nearby table.

a. Represent the training data in the 2D space.

b. What are the predictions of the 3- 5- and 7-nearest-neighbor classifiers at the point (1,1)?

x	y	
-1	1	$-$
0	1	$+$
0	2	$-$
1	-1	$-$
1	0	$+$
1	2	$+$
2	2	$-$
2	3	$+$



Solution:

b. $k = 3$: $+$; $k = 5$: $+$; $k = 7$: $-$.

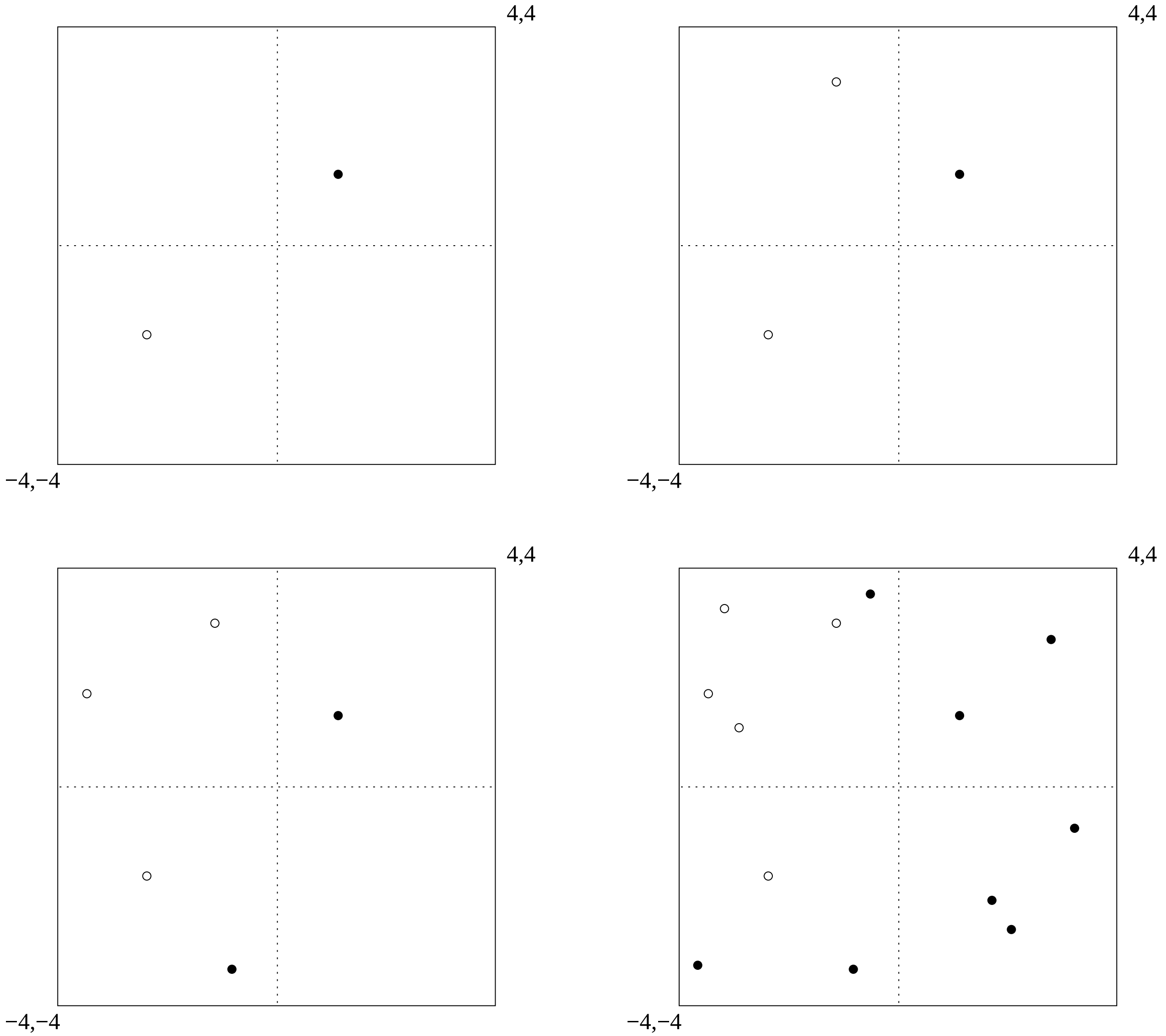
Drawing decision boundaries and decision surfaces for the 1-NN classifier

Voronoi Diagrams

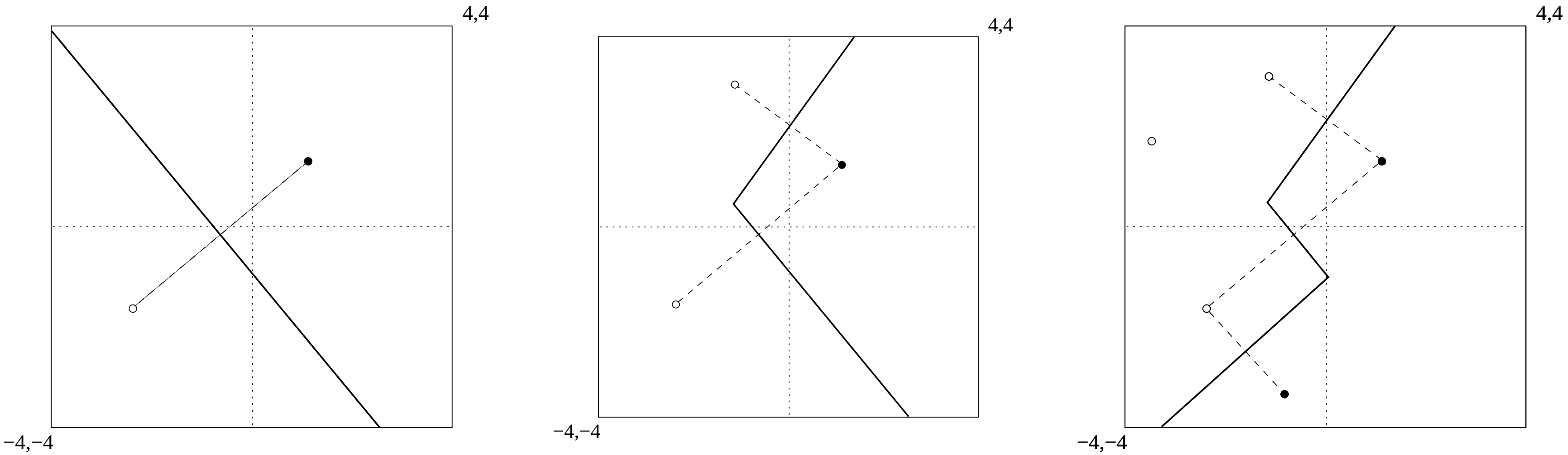
CMU, 2010 spring, E. Xing, T. Mitchell, A. Singh,
HW1, pr. 3.1

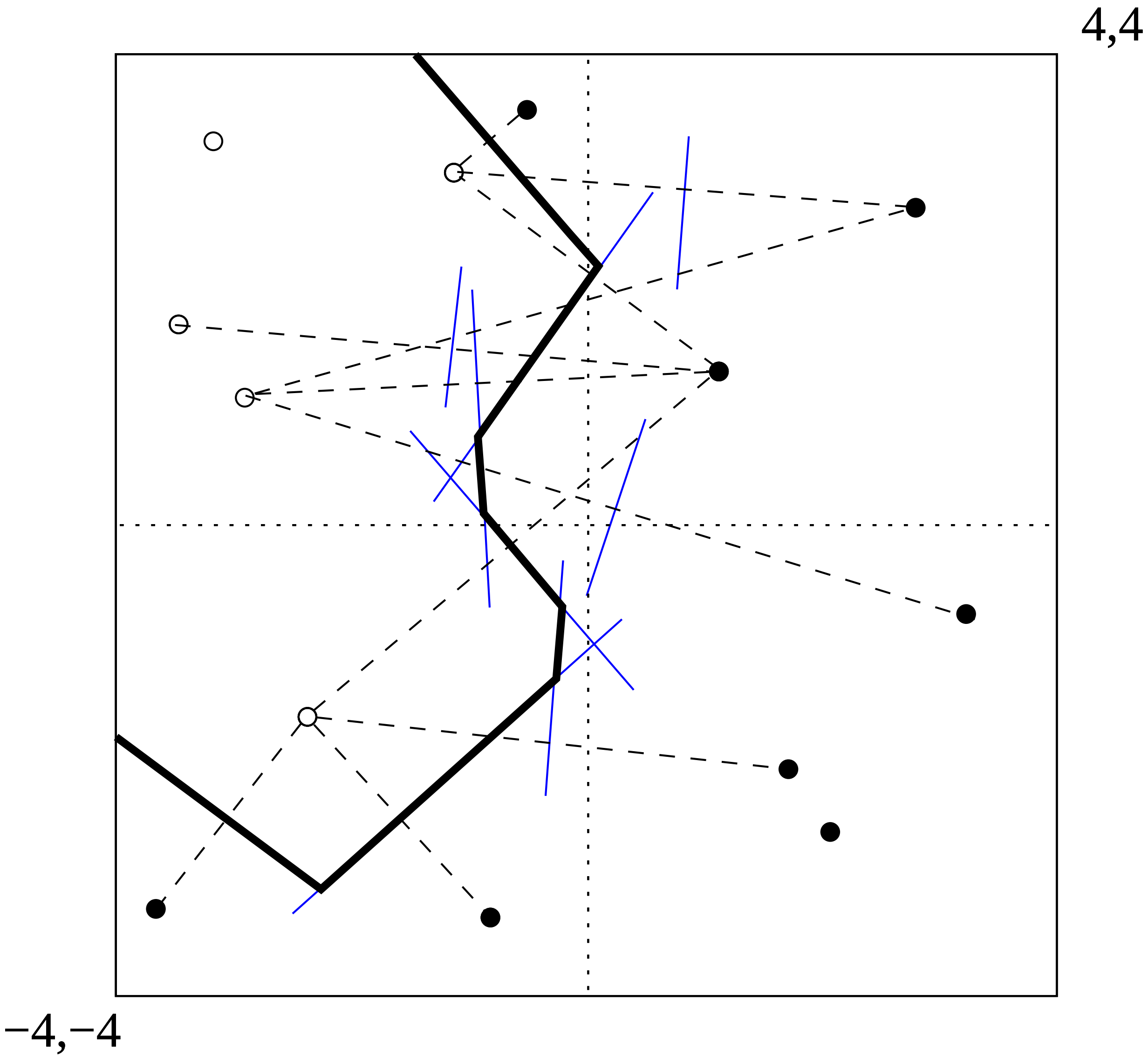
For each of these figures, we are given a few data points in 2-d space, each of which is labeled as either positive (blue) or negative (red).

Assuming that we are using the L2 distance as a distance metric, draw the decision boundary for the 1-NN classifier for each case.



Solution




4,4
−4,−4

Drawing decision boundaries and decision surfaces for the 1-NN classifier

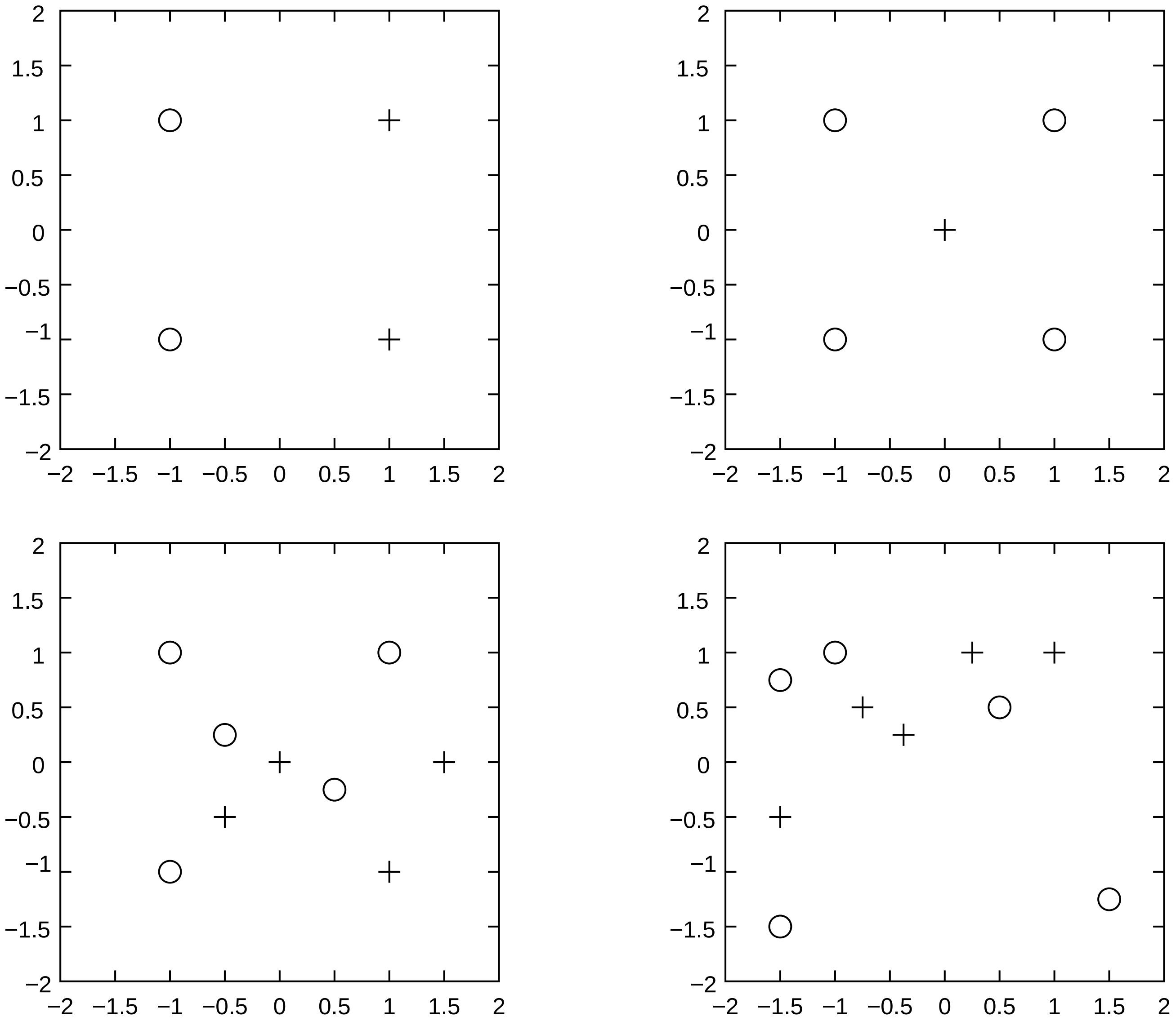
Voronoi Diagrams: DO IT YOURSELF

CMU, 2010 fall, Ziv Bar-Joseph, HW1, pr. 3.1

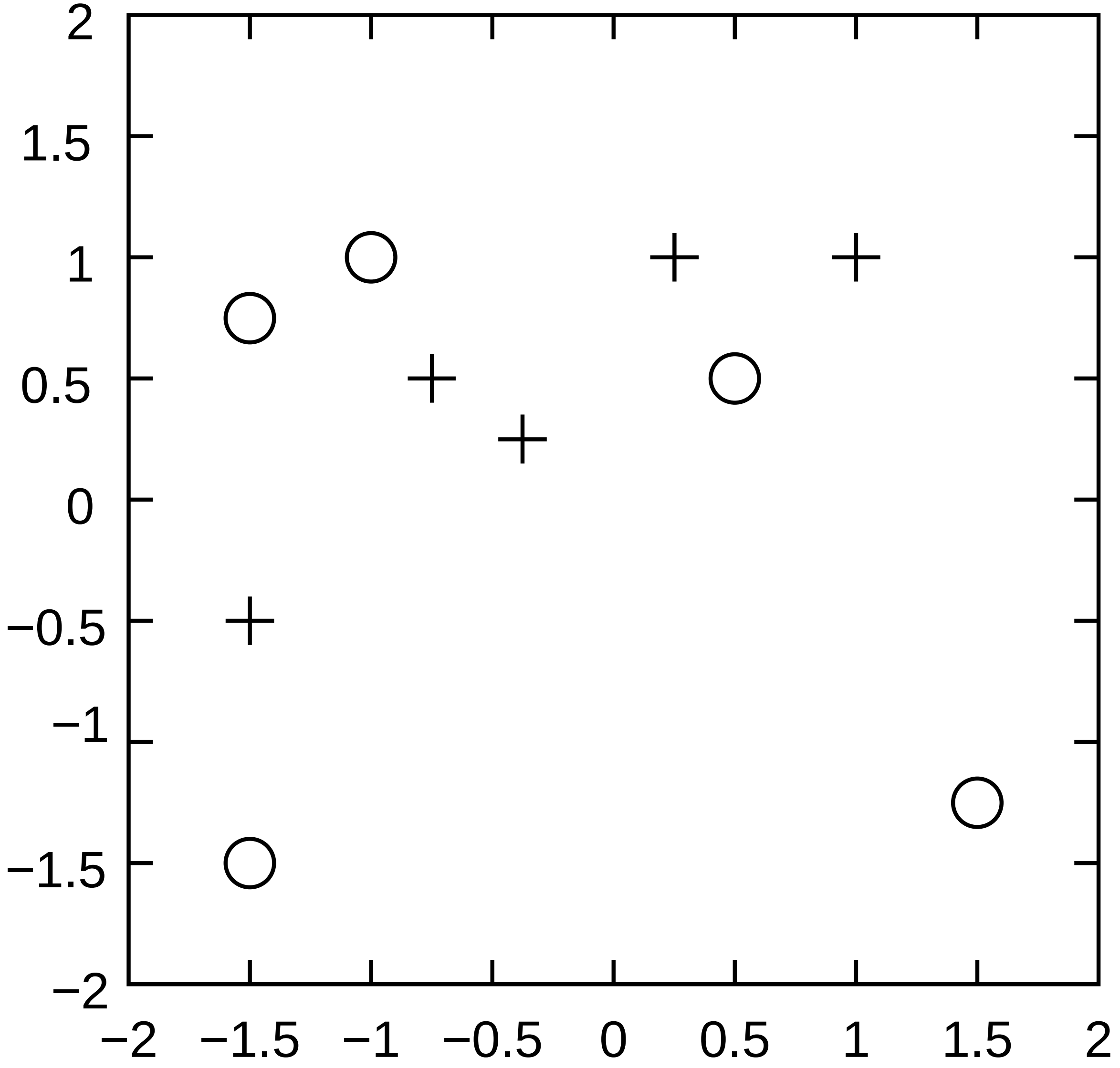
For each of the nearby figures, you are given negative (○) and positive (+) data points in the 2D space.

Remember that a 1-NN classifier classifies a point according to the class of its nearest neighbour.

Please draw the Voronoi diagram for a 1-NN classifier using Euclidean distance as the distance metric for each case.



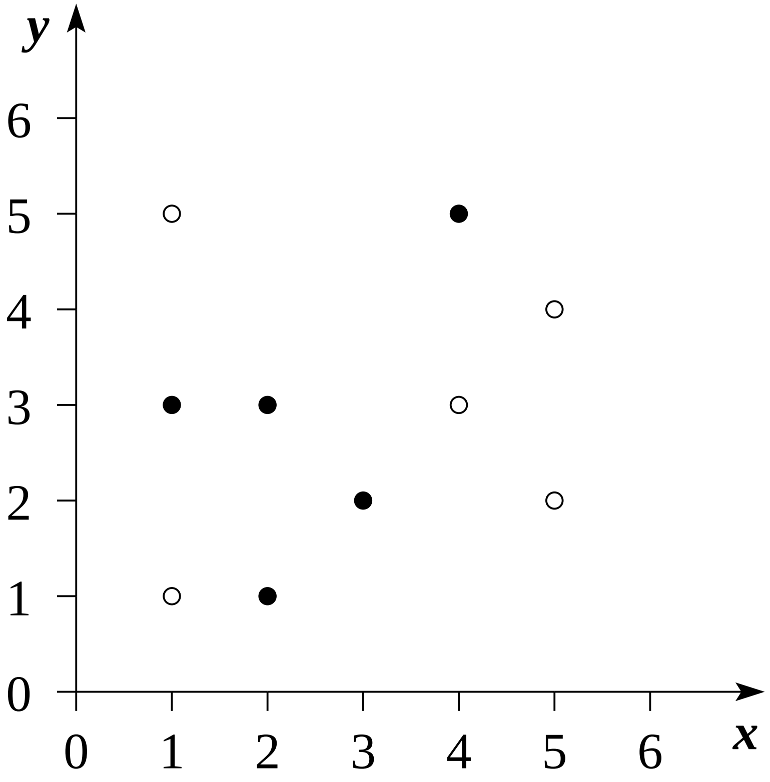
8.

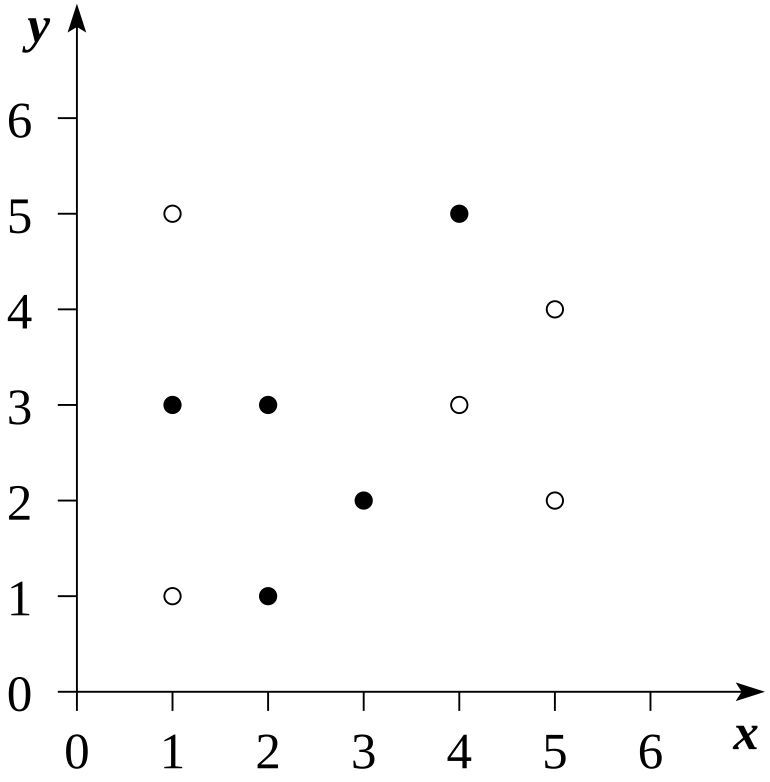


Decision boundaries and decision surfaces: Comparison between the 1-NN and ID3 classifiers

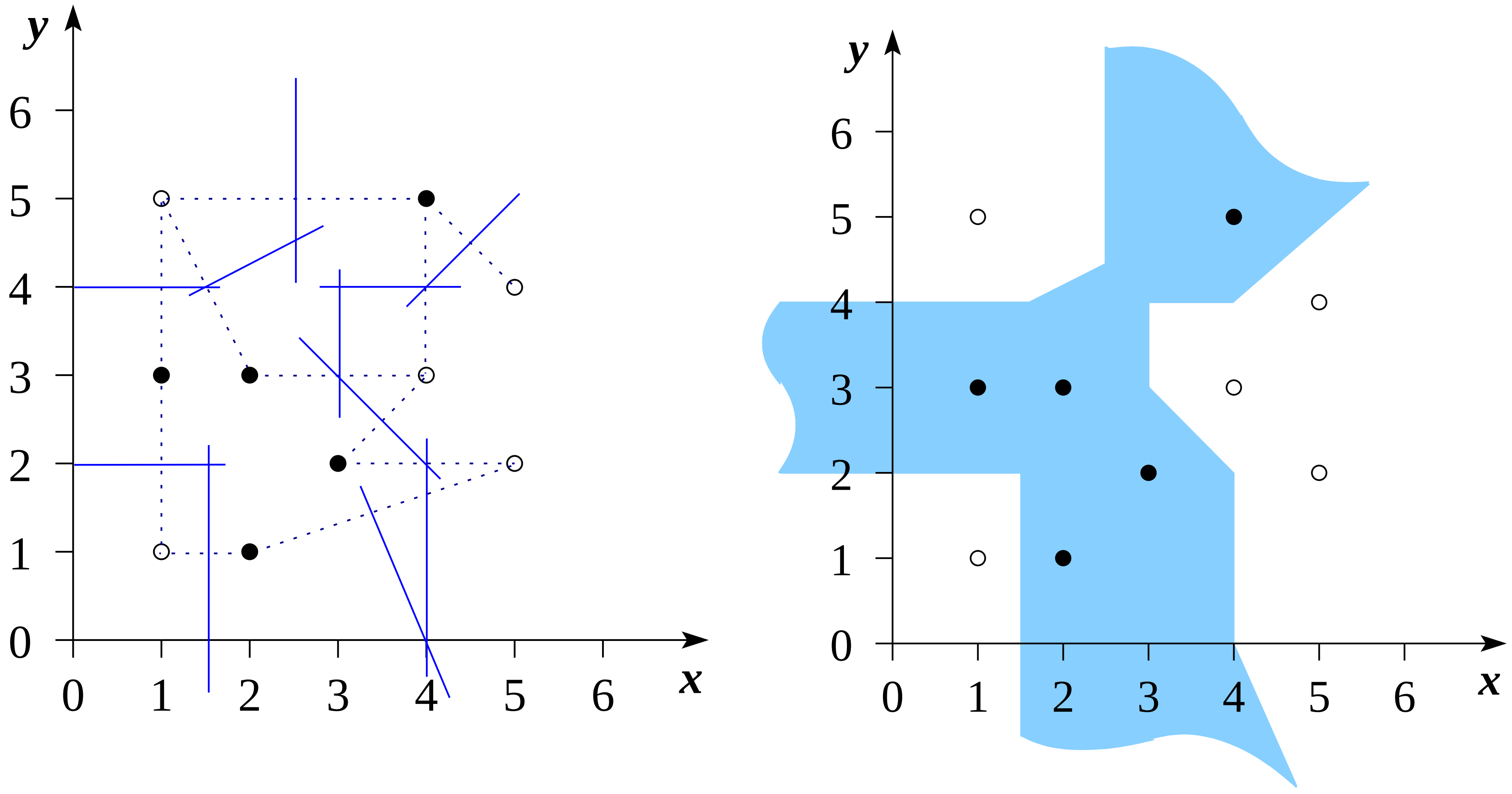
CMU, 2007 fall, Carlos Guestrin, HW2, pr. 1.4

For the data in the figure(s) below, sketch the decision surfaces obtained by applying
a. the K-Nearest Neighbors algorithm with $K = 1$;
b. the ID3 algorithm augmented with [the capacity to process] continous attributes.

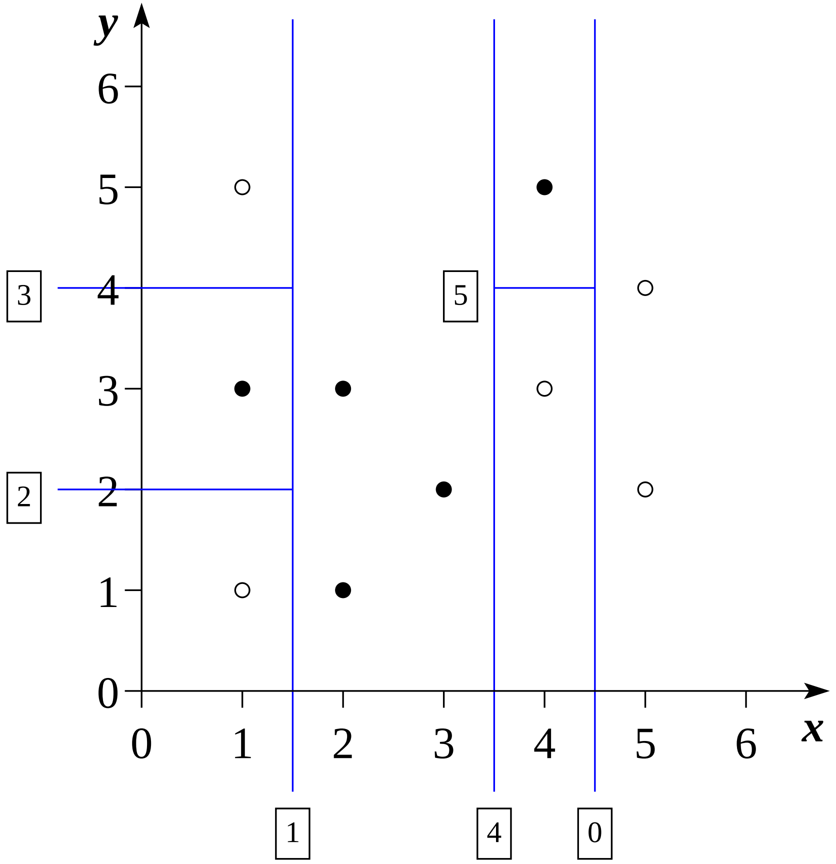


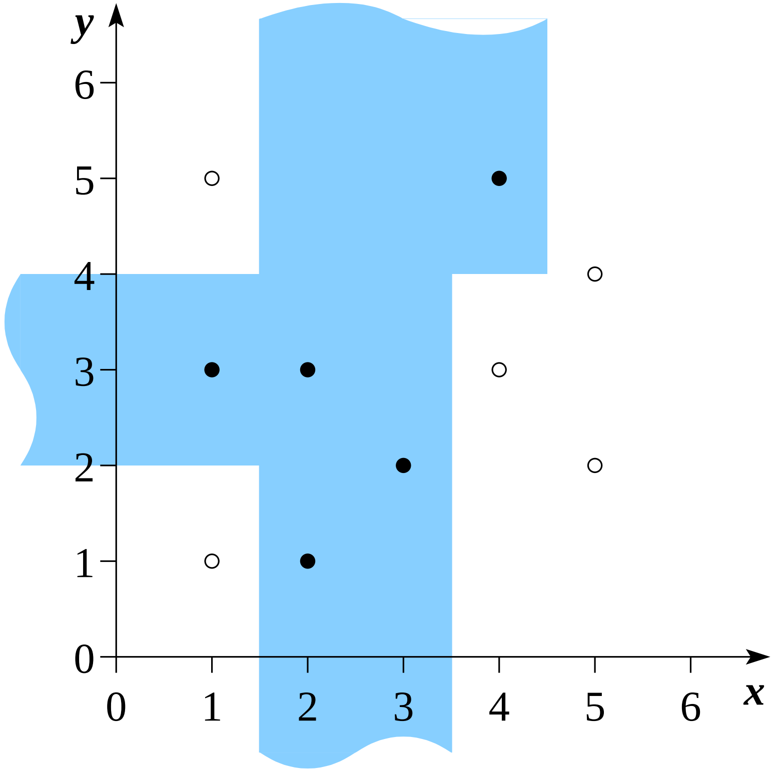


Solution: 1-NN



Solution: ID3





Instance-Based Learning
Some important properties

k-NN and the Curse of Dimensionality

Proving that the number of examples needed by k-NN grows exponentially with the number of features

CMU, 2010 fall, Aarti Singh, HW2, pr. 2.2

[Slides originally drawn by Diana Mînzat, MSc student, FII, 2015 spring]

Consider a set of n points $x_1, x_2, ..., x_n$ independently and uniformly drawn from a p-dimensional zero-centered unit ball

$$B = \{x \colon \|x\|^2 \leq 1\} \subset \mathbb{R}^p,$$

where $\|x\| = \sqrt{x \cdot x}$ and $\cdot$ is the inner product in $\mathbb{R}^p$.

In this problem we will study the size of the 1-nearest neighbourhood of the origin O and how it changes in relation to the dimension p, thereby gain intuition about the downside of k-NN in a high dimension space.

Formally, this size will be described as the distance from O to its nearest neighbour in the set $\{x_1, ..., x_n\}$, denoted by d^*:

$$d^* := \min_{1 \leq i \leq n} ||x_i||,$$

which is a random variable since the sample is random.

a. For $p = 1$, calculate $P(d^* \leq t)$, the *cumulative distribution function (c.d.f.)* of d^*, for $t \in [0, 1]$.

Solution:

In the one-dimensional space ($p = 1$), the unit ball is the interval $[-1, 1]$. The cumulative distribution function will have the following expression:

$$F_{n,1}(t) \overset{not.}{=} P(d^* \leq t) = 1 - P(d^* > t) = 1 - P(|x_i| > t, \text{ for } i = 1, 2, ..., n)$$

Because the points $x_1, ..., x_n$ were generated independently, the c.d.f. can also be written as:

$$F_{n,1}(t) = 1 - \prod_{i=1}^{n} P(|x_i| > t) = 1 - (1 - t)^n$$

b. Find the formula of the *cumulative distribution function* of d^* for the general case, when $p \in \{1, 2, 3, ...\}$.

Hint: You may find the following fact useful: the volume of a p-dimensional ball with radius r is

$$V_p(r) = \frac{(r\sqrt{\pi})^p}{\Gamma\left(\frac{p}{2} + 1\right)},$$

where Γ is Euler's Gamma function, defined by

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi},\ \Gamma(1) = 1,\ \textbf{and}\ \Gamma(x+1) = x\Gamma(x)\ \textbf{for any}\ x > 0.$$

Note: It can be easily shown that $\Gamma(n+1) = n!$ for all $n \in \mathbb{N}^*$, therefore the Gamma function is a generalization of the factorial function.

Solution:

In the general case, i.e. considering a fixed $p \in \mathbb{N}^*$, it is obvious that the cumulative distribution function of d^* will have a similar form to the $p = 1$ case:

$$\begin{aligned} F_{n,p}(t) \overset{not.}{=} P(d^* \leq t) &= 1 - P(d^* > t) = 1 - P(||\,x_i\,|| > t,\ i = 1, 2, \ldots, n) \\ &= 1 - \prod_{i=1}^{n} P(||\,x_i\,|| > t). \end{aligned}$$

Denoting the volume of the sphere of radius t by $V_p(t)$, and knowing that the points $x_1, ..., x_n$ follow a uniform distribution, we can rewrite the above formula as follows:

$$F_{n,p}(t) = 1 - \left(\frac{V_p(1) - V_p(t)}{V_p(1)}\right)^n = 1 - \left(1 - \frac{V_p(t)}{V_p(1)}\right)^n.$$

Using the suggested formula for the volume of the sphere, it follows immediately that $F_{n,p} = 1 - (1 - t^p)^n$.

c. What is the *median* of the random variable d^* (i.e., the value of t for which $P(d^* \leq t) = 1/2$) ? The answer should be a *function of* both the sample size n and the dimension p.

Fix $n = 100$ and plot the values of the median function for $p = 1, 2, 3, ..., 100$ with the median values on the y-axis and the values of p on the x-axis. What do you see?

Solution:

In order to find the median value of the random variable d^*, we will solve the equation $P(d^* \leq t) = 1/2$ of variable t:

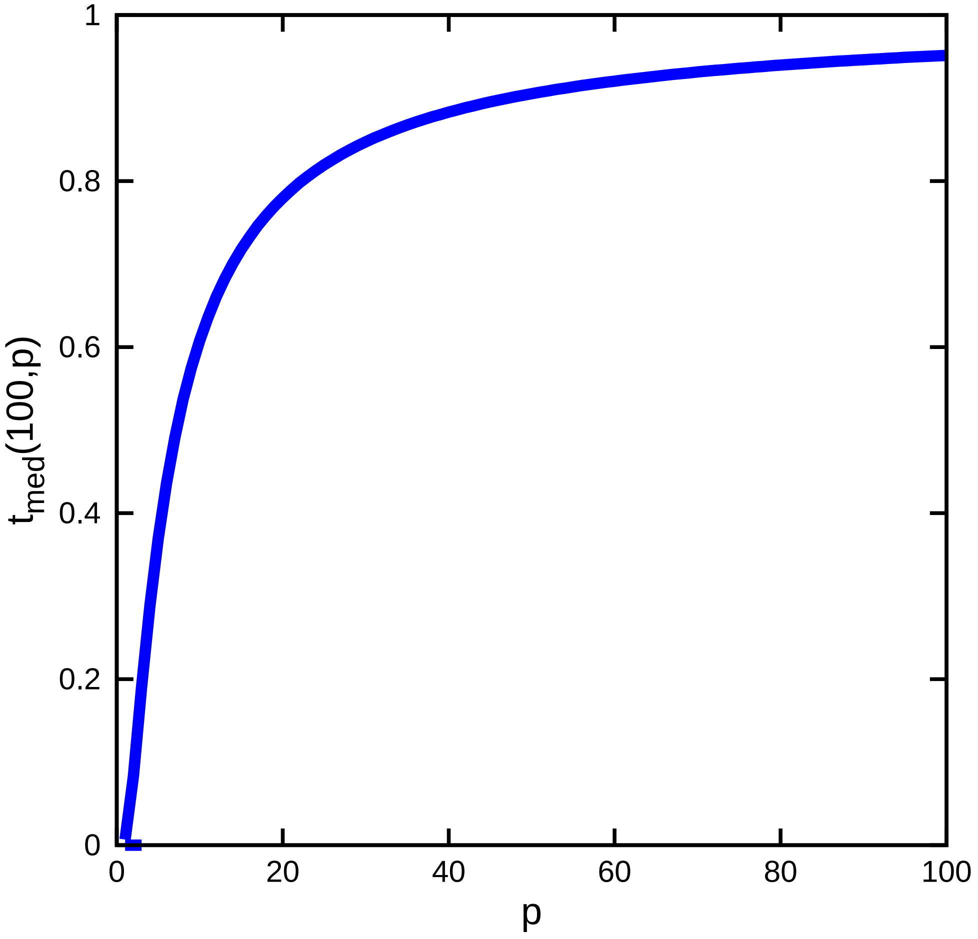
$$P(d^* \leq t) = \frac{1}{2} \;\Leftrightarrow\; F_{n,p}(t) = \frac{1}{2} \overset{b}{\Leftrightarrow} 1 - (1 - t^p)^n = \frac{1}{2} \Leftrightarrow (1 - t^p)^n = \frac{1}{2}$$

$$\Leftrightarrow\; 1 - t^p = \frac{1}{2^{1/n}} \Leftrightarrow t^p = 1 - \frac{1}{2^{1/n}}$$

Therefore,

$$t_{med}(n, p) = \left(1 - \frac{1}{2^{1/n}}\right)^{1/p}.$$

The plot of the function $t_{med}(100, p)$ for $p = 1, 2, \ldots, 100$:



Remark:

The minimal sphere containing the nearest neighbour of the origin in the set $\{x_1, x_2, ..., x_n\}$ grows very fast as the value of p increases.

When p becomes greater than 10, most of the 100 training instances are closer to the surface of the unit ball than to the origin O.

d. Use the c.d.f. derived at point b to determine how large should the sample size n be such that with probability at least 0.9, the distance d^* from O to its nearest neighbour is less than $1/2$, i.e., half way from O to the boundary of the ball.

The answer should be a *function* of p.
Plot this function for $p = 1, 2, \ldots, 20$ with the function values on the y-axis and values of p on the x-axis. What do you see?

***Hint*: You may find useful the Taylor series expansion of $\ln(1-x)$:**

$$\ln(1-x) = -\sum_{i=1}^{\infty} \frac{x^i}{i} \text{ for } -1 \leq x < 1.$$

Solution:

$$P(d^* \leq 0.5) \geq 0.9 \quad \Leftrightarrow \quad F_{n,p}(0.5) \geq \frac{9}{10} \overset{b.}{\Leftrightarrow} 1 - \left(1 - \frac{1}{2^p}\right)^n \geq \frac{9}{10} \;\Leftrightarrow\; \left(1 - \frac{1}{2^p}\right)^n \leq \frac{1}{10}$$

$$\Leftrightarrow \quad n \cdot \ln\left(1 - \frac{1}{2^p}\right) \leq \; -\ln 10 \Leftrightarrow n \geq \; \frac{\ln 10}{-\ln\left(1 - \dfrac{1}{2^p}\right)}$$

Using the decomposition of $\ln(1 - 1/2^p)$ **into a Taylor series (with** $x = 1/2^p$**), we obtain:**

$$\begin{aligned} & P(d^* \leq 0.5) \geq 0.9 \\ \Rightarrow \quad & n \geq (\ln 10)\; 2^p \frac{1}{1 + \frac{1}{2} \cdot \frac{1}{2^p} + \frac{1}{3} \cdot \frac{1}{2^{2p}} + \ldots + \frac{1}{n}\frac{1}{2^{(n-1)p}} + \ldots} \\ \Rightarrow \quad & n \geq \; 2^{p-1} \; \ln 10. \end{aligned}$$

***Note*:**

In order to obtain the last inequality in the above calculations, we considered the following two facts:

i. $\dfrac{1}{3 \cdot 2^p} < \dfrac{1}{4}$ **holds for any** $p \geq 1$**, and**

ii. $\dfrac{1}{n \cdot 2^{(n-1)p}} \leq \dfrac{1}{2^n} \Leftrightarrow 2^n \leq n \cdot 2^{(n-1)p}$ **holds for any** $p \geq 1$ **and** $n \geq 2$**.**

(This can be proven by induction on p**).**

So, we got:

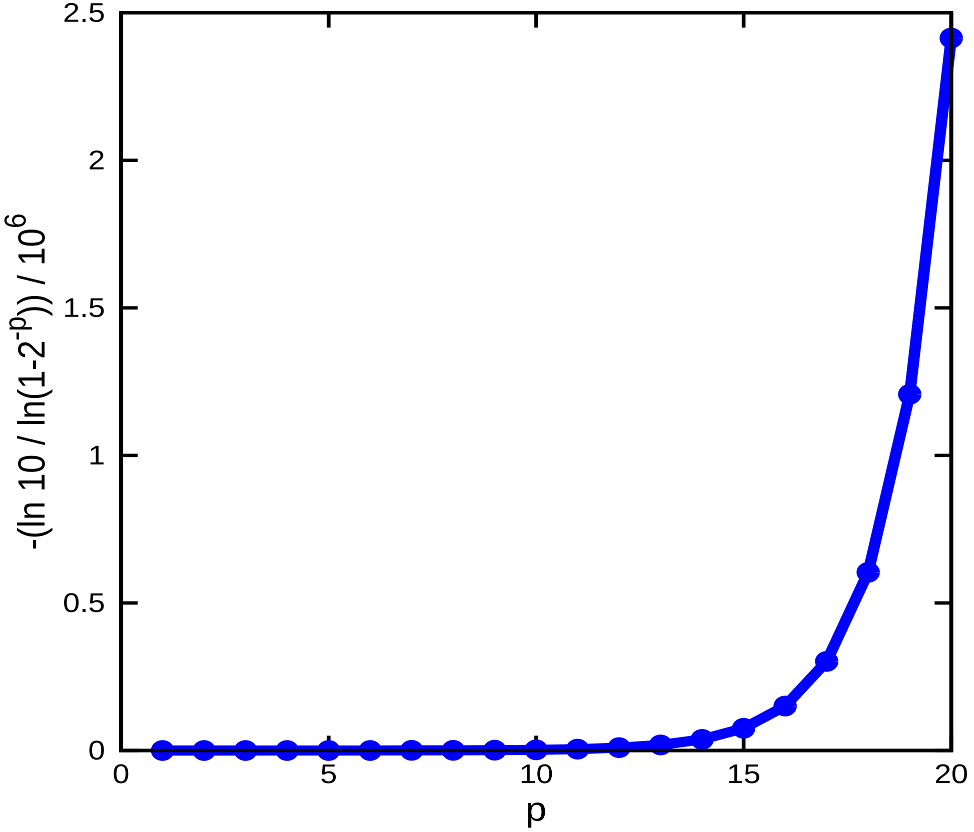
$$1 + \frac{1}{2} \cdot \frac{1}{2^p} + \frac{1}{3} \cdot \frac{1}{2^{2p}} + \ldots + \frac{1}{n} \frac{1}{2^{(n-1)p}} + \ldots <$$

$$1 + \frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n} + \ldots \rightarrow \frac{1}{1 - \frac{1}{2}} = 2.$$

The proven result

$$P(d^* \leq 0.5) \geq 0.9 \Rightarrow n \geq \ 2^{p-1} \ \ln 10$$

means that the sample size needed for the probability that $d^* < 0.5$ is large enough (9/10) grows exponentially with p.



e. Having solved the previous problems, what will you say about the downside of k-NN in terms of n and p?

Solution:

The k-NN classifier works well when a test instance has a "dense" neighbourhood in the training data.

However, the analysis here suggests that in order to provide a dense neighbourhood, the size of the training sample should be exponential in the dimension p, which is clearly infeasible for a large p.

(Remember that p is the dimension of the space we work in, i.e. the number of features of the training instances.)

An upper bound for the assimptotic error rate of 1-NN: twice the error rate of Joint Bayes

T. Cover and P. Hart (1967)

CMU, 2005 spring, Carlos Guestrin, HW3, pr. 1

Note: we will prove the *Covert & Hart' theorem* in the case of binary classification with real-values inputs.

Let $x_1, x_2, \ldots$ be the training examples in some fixed d-dimensional Euclidean space, and y_i be the corresponding binary class labels, $y_i \in \{0, 1\}$.

Let $p_y(x) \overset{not.}{=} P(X = x \mid Y = y)$ be the true conditional probability distribution for points in class y. We *assume* continuous and non-zero conditional probabilities: $0 < p_y(x) < 1$ for all x and y.

Let also $\theta \overset{not.}{=} P(Y = 1)$ be the probability that a random training example is in class 1. Again, *assume* $0 < \theta < 1$.

a. Calculate $q(x) \overset{not.}{=} p(Y=1 \mid X=x)$, the true probability that a data point x belongs to class 1. Express $q(x)$ in terms of $p_0(x), p_1(x)$, and θ.

Solution:

$$
\begin{aligned}
q(x) \quad & \overset{F.\ Bayes}{=} \quad \frac{P(X=x|Y=1)P(Y=1)}{P(X=x)} \\
& = \quad \frac{P(X=x|Y=1)P(Y=1)}{P(X=x|Y=1)P(Y=1)+P(X=x|Y=0)P(Y=0)} \\
& = \quad \frac{p_1(x)\,\theta}{p_1(x)\,\theta + p_0(x)(1-\theta)}
\end{aligned}
$$

b. The Joint Bayes classifier (usually called the Bayes Optimal classifier) always assigns a data point x the most probable class: $\text{argmax}_y\, P(Y = y \mid X = x)$.

Given some test data point x, what is the probability that example x will be misclassified using the Joint Bayes classifier, in terms of $q(x)$?

Solution:

The Joint Bayes classifier fails with probability $P(Y = 0|X = x)$ when $P(Y = 1|X = x) \geq P(Y = 0|X = x)$, and respectively with probability $P(Y = 1|X = x)$ when $P(Y = 0|X = x) \geq P(Y = 1|X = x)$. I.e.,

$$
\begin{aligned}
Error_{Bayes}(x) &= \min\{P(Y = 0|X = x), P(Y = 1|X = x)\} \\
&= \min\{1 - q(x),\, q(x)\} \\
&= \begin{cases} q(x) \ \textbf{if}\ q(x) \in [0\,,\, 1/2] \\ 1 - q(x) \ \textbf{if}\ q(x) \in (1/2\,,\, 1]. \end{cases}
\end{aligned}
$$

c. The 1-nearest neighbor classifier assigns a test data point x the label of the closest training point x'.

Given some test data point x and its nearest neighbor x', what is the *expected error* of the 1-nearest neighbor classifier, i.e., the probability that x will be misclassified, in terms of $q(x)$ and $q(x')$?

Solution:

$$\begin{aligned} Error_{1\text{-}NN}(x) &= P(Y=1|X=x)P(Y=0|X=x') + \\ & \quad\; P(Y=0|X=x)P(Y=1|X=x') \\ &= q(x)(1-q(x')) + (1-q(x))q(x'). \end{aligned}$$

d. In the asymptotic case, i.e. when the number of training examples of each class goes to infinity, and the training data fills the space in a dense fashion, the nearest neighbor x' of x has $q(x')$ converging to $q(x)$, i.e. $P(Y=1|X=x') \to p(Y=1|X=x)$.

(This is true due to *i.* the result obtained at the above point a, and *ii.* the assumed continuity of the function $p_y(x) \overset{not.}{=} p(X=x|Y=y)$ w.r.t. x.)

By performing this substitution in the expression obtained at point c, give the *asymptotic error* for the 1-nearest neighbor classifier at point x, in terms of $q(x)$.

Solution:

$$\lim_{x' \to x} Error_{\textit{1-NN}}(x) = 2q(x)(1-q(x))$$

e. Show that the asymptotic error obtained at point d is less than twice the Joint Bayes error obtained at point b and subsequently that this inequality leads to the corresponding relationship between the expected error rates:

$$E[\lim_{n\to\infty} Error_{1\text{-}NN}] \leq 2E[Error_{Bayes}].$$

Solution:

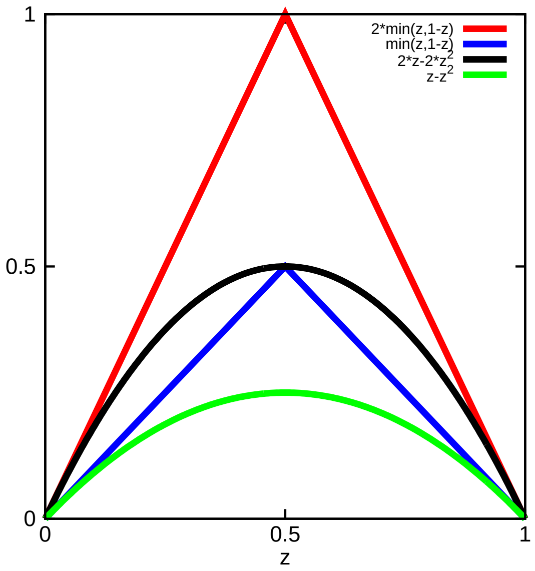
$z(1-z) \leq z$ for all z, in particular for $z \in [0\,,\,1/2]$, and $z(1-z) \leq 1-z$ for all z, in particular for $z \in [1/2\,,\,1]$. Therefore, for all x,

$$q(x)(1-q(x)) \leq \begin{cases} q(x) \textbf{ if } q(x) \in [0\,,\,1/2] \\ 1-q(x) \textbf{ if } q(x) \in (1/2\,,\,1]. \end{cases}$$

The results obtained at points b and d lead to

$$\lim_{n\to\infty} Error_{1\text{-}NN}(x) = 2q(x)(1-q(x)) \leq 2Error_{Bayes}(x) \textbf{ for all } x.$$

By multiplication with $P(x)$ and suming upon all values of x, we get: $E[\lim_{n\to\infty} Error_{1\text{-}NN}] \leq 2E[Error_{Bayes}]$.



Remarks

- $E[\lim_{n\to\infty} Error_{1\text{-}NN}] \geq E[Error_{Bayes}]$

  Proof:
  $2z - 2z^2 \geq z \;\forall z \in [0\,,\,1/2]$ **and** $2z - 2z^2 \geq 1 - z \;\forall z \in [1/2\,,\,1]$**.**
  Therefore,

  $$2q(x)(1 - q(x)) \geq Error_{Bayes}(x) \textbf{ for all } x,$$

  and

  $$\lim_{n\to\infty} Error_{1\text{-}NN}(x) = \lim_{x'\to x} Error_{1\text{-}NN}(x) \geq Error_{Bayes}(x) \textbf{ for all } x.$$

- **The Cover & Hart' upper bound for the asymptotic error rate of 1-NN doesn't hold in the non-asymptotic case (where the number of training examples is finite).**

Remark

from: ***An Elementary Introduction to Statistical Learning Theory*****,**
S. Kulkarni, G. Harman, 2011, pp. 68-69

An even tighter upper bound exists for $E[\lim_{n \to \infty} Error_{1\text{-}NN}]$**:**
$2E[Error_{Bayes}](1 - E[Error_{Bayes}])$

Proof:

From $\lim_{x' \to x} Error_{1\text{-}NN}(x) = 2q(x)(1 - q(x))$ **(see point** d**) and**
$Error_{Bayes}(x) = \min\{1 - q(x),\ q(x)\}$ **(see point** b**),**

it follows that
$\lim_{x' \to x} Error_{1\text{-}NN}(x) = 2Error_{Bayes}(x)(1 - Error_{Bayes}(x))$**.**

By multiplying this last equality with $P(x)$ **and suming on all** x **— in fact, integrating upon** x **—, we get**

$$E[\lim_{x' \to x} Error_{1\text{-}NN}] = 2E[Error_{Bayes}(1 - Error_{Bayes})] = 2E[Error_{Bayes}] - 2E[(Error_{Bayes})^2].$$

Since $E[Z^2] \geq (E[Z])^2$ **for any** Z **(** $Var(Z) \overset{def.}{=} E[(Z - E[Z])^2] \overset{comp.}{=} E[Z^2] - (E[Z])^2 \geq 0$**),**
it follows that

$$E[\lim_{x' \to x} Error_{1\text{-}NN}] \leq 2E[Error_{Bayes}] - 2(E[Error_{Bayes}])^2 = 2E[Error_{Bayes}](1 - E[Error_{Bayes}]).$$

Other Results

[from *An Elementary Introduction to Statistical Learning Theory*,
S. Kulkarni, G. Harman, 2011, pp. 69-70]

- **When certain restrictions hold,**

$$E[\lim_{n\to\infty} Error_{k\text{-}NN}] \leq \left(1+\frac{1}{k}\right) E[Error_{Bayes}].$$

○ **However, it can be shown that there are some distributions for which 1-NN outperforms k-NN for any fixed $k > 1$.**

- **If $\dfrac{k_n}{n} \to 0$ for $n \to \infty$ (for instance, $k_n = \sqrt{n}$), then**

$$E[\lim_{n\to\infty} Error_{k_n\text{-}NN}] = E[Error_{Bayes}].$$

Significance

The last result means that k_n-NN is

- a ***universally consistent learner*** (because when the amount of training data grows, its performance approaches that of Joint Bayes) and
- ***non-parametric*** (i.e., the underlying distribution of data can be arbitrary and we need no knowledge of its form).

Some other universally consistent learners exist.

However, the ***convergence rate*** is critical. For most learning methods, the convergence rate is very slow in high-dimensional spaces (due to "the curse of dimensionality"). It can be shown that ***there is no "universal" convergence rate,*** i.e. one can always find distributions for which the convergence rate is arbitrarily slow.

There is no one learning method which can universally beat out all other learning methods.

Conclusion

Such results make the ML field continue to be exciting, and makes the design of good learning algorithms and the understanding of their performance an important science and art!

On 1-NN and kernelization with RBF

CMU, 2003 fall, T. Mitchell, A. Moore, final exam, pr. 7.f

After mapped into the feature space Q through a radial basis kernel function (RBF), the 1-NN algorithm using unweighted Euclidean distance may be able to achieve a better classification performance than in the original space (though we can't guarantee this). True or False?

Answer

Consider $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^n$ such that $K(x,y) \overset{not.}{=} e^{-\frac{||x-y||^2}{2\sigma^2}} = \phi(x)\cdot\phi(y), \forall x,y \in \mathbb{R}^d$. $\mathbb{R}^d$ is the original space, $\mathbb{R}^n$ is the "feature" space, and $e^{-\frac{||x-y||^2}{2\sigma^2}}$ is the radial basis function (RBF). Then

$$\begin{aligned} & ||\phi(x)-\phi(y)||^2 = (\phi(x)-\phi(y))\cdot(\phi(x)-\phi(y)) \\ = \quad & \phi(x)\cdot\phi(x)+\phi(y)\cdot\phi(y)-2\cdot\phi(x)\cdot\phi(y) = e^{-\frac{||x-x||^2}{2\sigma^2}} + e^{-\frac{||y-y||^2}{2\sigma^2}} - 2\cdot e^{-\frac{||x-y||^2}{2\sigma^2}} \\ = \quad & e^0 + e^0 - 2\cdot e^{-\frac{||x-y||^2}{2\sigma^2}} = 2 - 2\cdot e^{-\frac{||x-y||^2}{2\sigma^2}} = 2 - K(x,y) \end{aligned}$$

Suppose x_i and x_j are two neighbors for the test instance x such that $\|x-x_i\| < \|x-x_j\|$. After mapped to the feature space,

$$\|\phi(x)-\phi(x_i)\|^2 < \|\phi(x)-\phi(x_j)\|^2 \Leftrightarrow 2-K(x,x_i) < 2-K(x,x_j) \Leftrightarrow K(x,x_i) > K(x,x_j)$$

$$\Leftrightarrow e^{-\frac{||x-x_i||^2}{2\sigma^2}} > e^{-\frac{||x-x_j||^2}{2\sigma^2}} \Leftrightarrow -\frac{||x-x_i||^2}{2\sigma^2} > -\frac{||x-x_j||^2}{2\sigma^2} \Leftrightarrow ||x-x_i||^2 < ||x-x_j||^2.$$

So, if x_i is the nearest neighbor of x in the original space, it will also be the nearest neighbor in the feature space. Therefore, 1-NN doesn't work better in the feature space. (The same is true for k-NN.)

***Note*: k-NN using non-Euclidean distance or weighted voting may work.**