Nearest Neighbor Pattern Classification

T. M. Cover and P. E. Hart

May 15, 2018

1 The Intro

The nearest neighbor algorithm/rule (NN) is the simplest nonparametric decisions procedure, that assigns to unclassified observation (incoming test sample) the class/category/label of the nearest sample (using metric) in training set .

In this paper we shall show that in large sample case or number of training set approach infinity , the probability of error to NN is bounded by (lower bound) R^* Bayes probability of error of classification and by upper bound $2R^*(1-R^*)$

2 The Nearest Neighbor Rule

A set of n pairs is given $(x_1, \theta_2), ...(x_n, \theta_n)$ s.t the $x_i's$ take value in a metric space X upon which is defined a metric d.

The category θ_i is assigned to the ith sample or individual from finite subset {1,2, ... M} and x_i is the measurement made upon that ith.

we shall say "x_i belongs to θ_i" when we mean precisely that the ith sample have measurement x_i with category θ_i .

Our Goal is to classify the measurement x , for a new arriving pair (x, θ), the classification is procedure to estimate the θ since x is observable .

The estimation of θ is done by utilizing the fact that we have set of n correctly classified points.

We shall call :

$$x_n^{'} \in (x_1, x_2, x_3, ...x_n)$$

nearest neighbor of x if

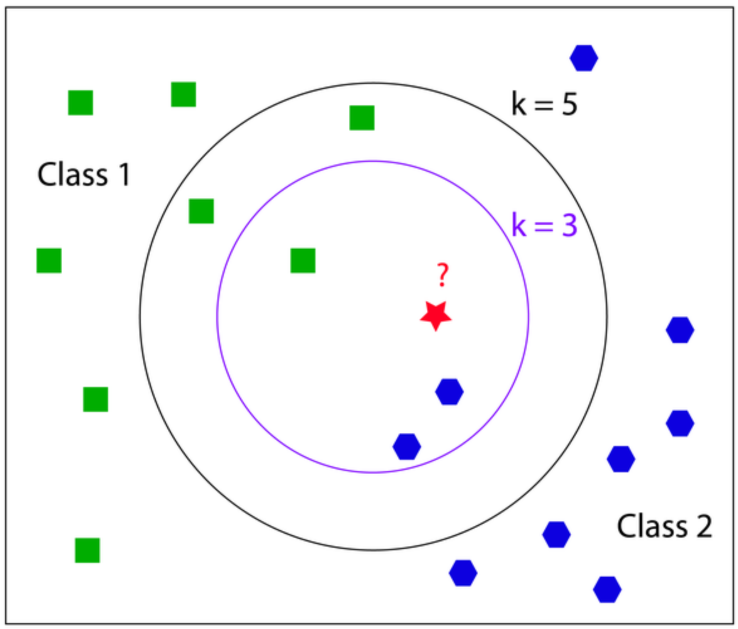
$$min\ d(x_i, x) = d(x_n^{'}, x)\ i = 1, 2, 3..., n.$$

In this case the decision of the nearest neighbor rule is assigning the category $\theta_n^{'}$ to our new measurement x .

The 1-NN rule decide x belongs to the category of nearest neighbor and ignored the others!.

In general k-NN rule decide x belongs to the category of majority vote of the nearest k neighbors.

Example : Given M=2 (i.e 2 categories) ,green squares , blue hexagon,a new arriving measurement (i.e the red star) and we desire to classify it.



1-NN rule : it will classified as blue hexagon.
3-NN rule : it will classified as blue hexagon.
5-NN rule : it will classified as green square.

3 The Admissibility Of Nearest Neighbor Rule

If we have large number of samples it makes good sense to use instead of classifying based on the nearest neighbor (1-NN) , we can use the idea of classifying depend on the majority vote of the nearest K neighbors(K-NN).

(What is the problem using K very large , What is the problem using small K?).

(*)picking a very large K will decrease the effect of nearest or distance factor s.t the label will assigned to a new sample depend on whole system.

If we take K equal to number of samples , then each new coming sample will labeled depend on the majority of exist labeled samples. (We can solve with "weighted kNN).

(*)picking a very small K will let outliers affect the label system.

(What is the problem using K even ?).

The purpose of this section show that among the class the k-NN rules, single nearest neighbor rule (1-NN) is admissible because th single-NN has strictly lower probability than any other k-NN.

Example :
(*)The prior probability $\eta_1 = \eta_2 = 1/2$, the probability of choosing the densities functions.
(*)The conditional densities f_1, f_2 are uniform over unit disk D_1, D_2 with centers $(-3, 0), (3, 0)$.

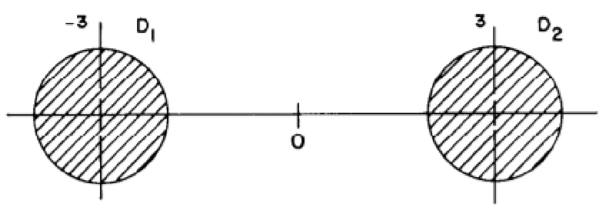


Figure 1: Admissibility of nearest neighbor rule.

The probability that j individuals from n samples come from category 1 , and hence have measurements lying in D_1 is :
$(1/2)^n * \binom{n}{j}$

Without loss of generality ,assume that the unclassified x lies in category 1 , Then the 1-NN make a classification error only if the nearest neighbor $x_n^{'}$ belongs to category 2 and thus lies in D_2. But from inspection of the distance relationships , if the nearest neighbor to x in D_2 , then each x_i must lie in D_2 .

Why ?

1 - if we assume having at least one $x_i's$ from n-samples in D_1 , then 1-NN will not make classification error, because it will classify our x according to the assumption of x that lies in D_1 , because these $x_i's$ one of them is the nearest. (as Shown in Fig.2)

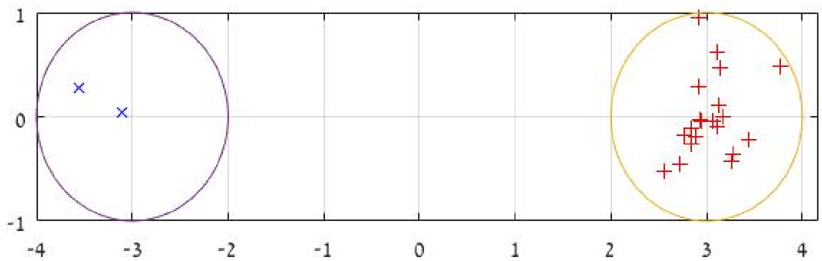


Figure 2: x in D_1 with another sample
1-NN rule will classify x as 1

2- making a classification error we need the compliment assumption s.t non-of the samples or $x_i's$ are exist in D_1 , then 1-NN will classify the x according to nearest which exist in D_2.(as Shown in Fig.3). to get such error we need that all individuals came from category 2. Thus the probability $P_e(1;n)$ of error of the NN rule in this case is precisely $(1/2)^n$, which mean the probability that all $x_1, x_2, ...x_n$ all lie in D_2.

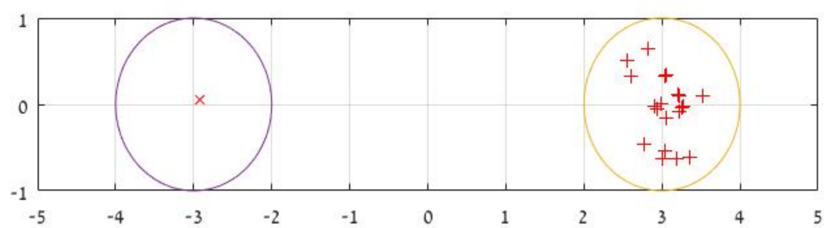


Figure 3: x is D_1
1-NN rule will classify x as 2

in general Let $k = 2k_0 + 1$, k is odd number,then k-NN rule make error if k_0 or fewer points lie in D_1 ,This occurs with probability.:

$P_e(k;n) = (1/2)^n * \sum_{j=0}^{k_0} \binom{n}{j}$. (why ?)

The explanation same as above because k is odd number and k_0 is smaller than half of k, and k-NN rule depend on majority vote then if there are k_0 or fewer points in D_1 it will give error in classification .

as we see that the 1-NN rule has strictly lower P_e than any of other k-NN rule, the probability of error increase when k is increased , then we prefer to choose 1-NN (as Shown in Fig.4)

In last we have some points :

1- $P_e(k;n) \uparrow 1/2$ in k , for any n.(why ?) (as Shown in Fig.4)

2- $P_e(k;n) \downarrow 0$ in n , for any k > 0. (as Shown in Fig.5)

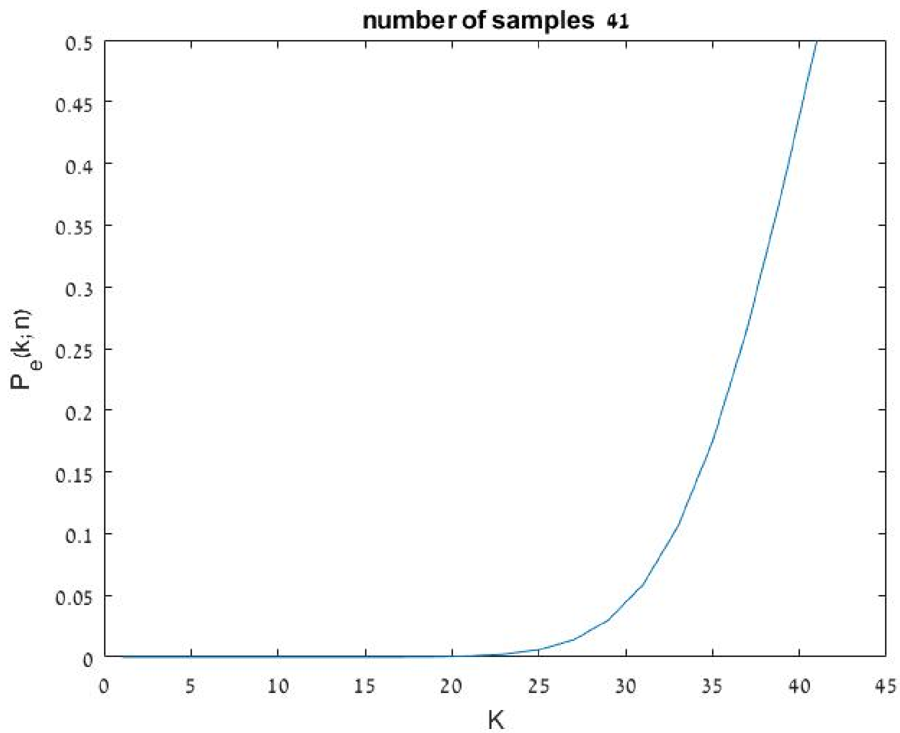


Figure 4: $P_e(k; n)$ as function of k
P_e is monotonically increasing and bounded by 1/2 in worst case , which mean pure guessing.

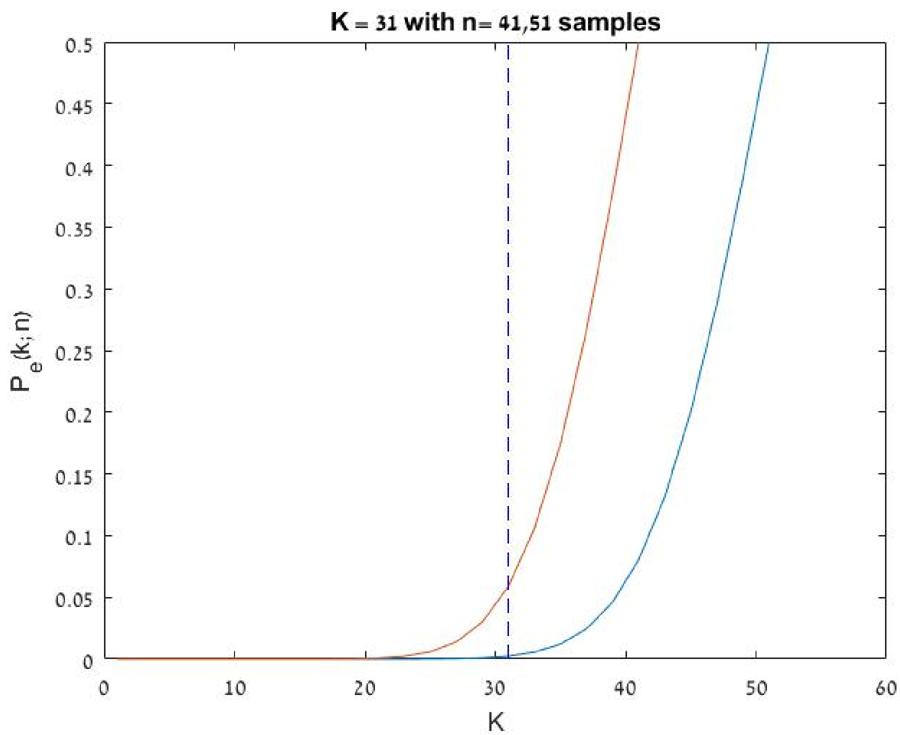


Figure 5: $P_e(k;n)$ as function of k
P_e for specific K the error function P_e is decreasing when n is increasing

3- $P_e(k_n; n) \to 0$ if $0 < k_n/n \leq \alpha < 1$ for all n. The limit of $P_e(k_n; n)$ is zero when the n approach infinite or big numbers of samples .(why ? proof derived from definition of $P_e(k_n; n)$).

In general then 1-NN rule is strictly better than other version of k-NN in those cases where the supports of the densities $f_1, f_2, ...f_M$ are such that each in-class have distance that greater than any between-class distance .

4 Bayes Procedure

In Bayesian Decision Theory the Basic Idea to o minimize errors, is choosing the least risky class, i.e. the class for which the expected loss is smallest. minimizing the probability of error in classifying a given observation x into one of M categories/classes , and all the statistic will be assumed known .

Let x denote the measurements on an individual and X the sample space of possible values of x , we shall refer to x as observation. we aim to build a decision for x to one of the M categories/classes which give us the least probability error.

For the purposes of defining the Bayes risk :

4.1 2-class Bayes decision : sea-bass and salmon

1 - Let $\eta_1, \eta_2, ...\eta_M$, $\eta_i \geq 0$, $\sum_{i=1}^{M} \eta_i = 1$ prior probability of the M categories.
.

For our problem : Let $\omega_1 = sea - bass, \omega_2 = salmon$.
$P(\omega_1)$ is the probability to get sea-bass.
$P(\omega_2)$ is the probability to get salmon..
$P(\omega_1) + P(\omega_2) = 1$.

2 - We assume class(category)-conditional probability density function(p.d.f) are known. $P(x|\theta_i) = f_i(x)$: is the probability of x to be category i , it's calculate the likelihood of observation x being in class i . (as Shown in Fig.6)

For our problem : we can take any property , for example we take the weight as measurements (x). in this case the likelihood, "what is the probability of observed/measured weight (x) for given fish type."

3- The posterior probability , posterior probability for a class/category: the probability of a class given the observation.

For our problem : how much the observed weight describe the type of fish .!! "What is the probability of class for given weight (x)."

Let $\eta_i \bar{(}x) = P(\theta_i|x)$ the posterior probability ,by the Bayes theorem :

Posterior $= \dfrac{likelihood * prior}{evidence}$.

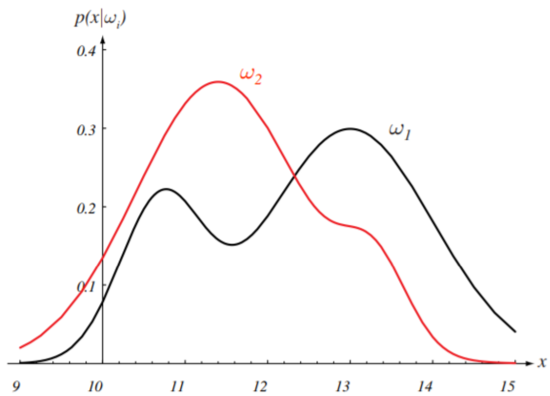


Figure 6: conditional p.d.f over observation x

$$\bar{\eta_i}(x) = \frac{\eta_i * f_i(x)}{\sum_{i=1}^{M} \eta_i * f_i(x)}$$

/*the denominator using Law of total probability*/

Example : (as show in Figure 6) x = 9 give very strong indication for ω_2 than ω_2 ,while intersection between 2 graphs don't give any indication about type of class.

Example : ω_1, ω_2 are two classes with prior probability $2/3, 1/3$ respectively.

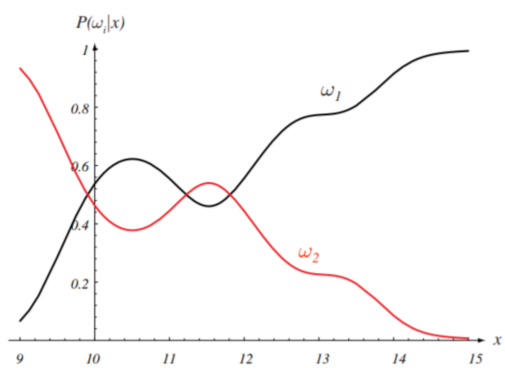


Figure 7: Posterior Probability

These graph calculated using likelihood graph in (as Shown in Fig.7) and the prior probability and evidence.

Let $L(i, j)$ be the loss incurred by assigning an individual from category i to category j. The cost associated with taking decision j with being i the correct category/class.

"What is the loss if we choose for observed x is salmon while in real it's sea-bass (or vice-versa)."

if the statistician decide to place an individual with measurement x (observation) into category j , then the conditional loss/risk or the expected loss is :

$r_j(x) = \sum_{i=1}^{M} \bar{\eta}_i(x) * L(i,j)$.

We have all these decision $r_1(x), r_2(x), ...r_m(x)$ every decision have its own loss.

In our case $r_{salmon}(x)$ or $r_{sea-bass}(x)$.

For a given x the conditional loss is minimum when the observation or individual is assigned to the category j, $r_j(x)$ is the lowest among all other decisions. Bayes decision rule is given by deciding the category j for which $r_j(x)$ is the lowest ,which mean we drop from the loss conditional is $\bar{\eta}_i(x) * L(i,j)$ the biggest value and we remain with small loss.

the conditional(for specific observation x) Bayes risk is :

$$r^*(x) = min_i \left\{ \sum_{i=1}^{M} \bar{\eta}_i(x) * L(i,j). \right\}$$

.

Overall risk : Suppose we have function $\alpha(x)$ that determine for each individual x a general decision rule (i.e a category 1,...M).

$r_{\alpha(x)}$ (the conditional risk/loss with respect to α).

The overall risk or the expectation of loss/risk with respect to α is defined as:

$R_{\alpha(x)} = E[r_{\alpha(x)}] = \int r_{\alpha(x)}(x) * p(x)dx$.

Overall Bayes risk: instead of using random function to determine the category , if we use the function that return lowest loss/risk (Bayes rule) we get Bayes risk and it's the expectation of loss with respect to Bayes rule decision $r^*(x)$ we get optimal or best results.

$minR = R^* = \int r^*(x) * p(x)dx = E[r^*(x)]$.

5 Convergence Of Nearest Neighbors

We first prove that the nearest neighbor of x converges almost to x as the training size grows to infinity.

Theorem Convergence of 1-NN if $x, x_1, x_2, ..., x_n$ are i.i.d set in a separable metric space X, $x_n^{'}$ is the nearest neighbor to x .

Proof :

Let $S_x(r)$ be the sphere centered at x, of radius r. $\bar{x} \in X : r \geq d(x, \bar{x})$, d is metric defined of X .

** The probability that nearest neighbor if x does not fall in the $S_x(\delta)$, $\delta > 0$ is the probability that no point in the training set fall within such sphere .

Probability of point to falls inside is the sphere is $\int_{x' \in S_x(\delta)} p(x^{'}) dx'$ -¿ sum over all points that in sphere.

$Pr\{d(x_n^{'}, n) > \delta\} = Pr\{x_n^{'} \notin S_x(\delta)\} = (1 - Pr\{S_x(\delta)\})^n \to 0$

We can conclude that this property will more hard to be satisfied when our training set size are going to infinity

6 Nearest Neighbors and Bayes Risk

Let $x_n^{'} \in x_1, x_2, ..x_n$ be the nearest to let $\theta_n^{'}$ be the class/category that belong to the individual $x_n^{'}$

The loss of assigning θ as $\theta_n^{'}$ is $L(\theta, \theta_n^{'})$

We define the n-sample NN risk $R(n)$ by the expectation of loss

$$R(n) = E[L(\theta, \theta_n^{'})]$$

and the in large sample R = $\lim_{n \to \infty} R(n)$

6.1 M = 2

2-category 0,1 binary classification problem the probably of error criterion given by 0-1 matrix loss.

$$L = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

L counts an error whenever a mistake in classification is made .

Theorem : 2-category

Let X be a separable metric space , f_1, f_2 be likelihood probability with sum equal to one $f_1(x) + f_2(x) = 1$, Then the NN risk R (probability of error) has the bounds

$$2R^*(1 - R^*) \geq R \geq R^*$$

Proof : Let us take the random variable x and $x_n^{'}$ in the n-samples NN problem , The conditional NN risk (conditional expectation) $r(x, x_n^{'})$ is given upon $\theta, \theta_n^{'}$ by :

$r(x, x_n^{'}) = [L(\theta, \theta_n^{'})|x, x_n^{'}] = Pr\{\theta \neq \theta_n^{'}|x, x_n^{'}\} = Pr\{\theta = 1|x\}Pr\{\theta_n^{'} = 2|x_n^{'}\} + Pr\{\theta = 2|x\}Pr\{\theta_n^{'} = 1|x_n^{'}\}$

As we see that $Pr\{\theta = 1|x\}$ is the posterior probability denoted by $\bar{\eta_1}(x)$ then we can rewrite equation above :

$$r(x, x_n^{'}) = \bar{\eta_1}(x) * \bar{\eta_2}(x_n') + \bar{\eta_1}(x_n') * \bar{\eta_2}(x)$$

By the lemma $x_n^{'}$ convergence to the random variable x with probability one .Hence with probability one,

$$\bar{\eta(x_n')} \rightarrow \bar{\eta(x)}$$

then the conditional risk , with probability one can be written :

$$r(x, x_n^{'}) \rightarrow r(x) = 2 * \bar{\eta_1(x)} * \bar{\eta_2(x)}$$

where r(x) is the limit of to n-sample conditional NN risk;
the conditional Bayes risk is:

$$r^*(x) = min\{\bar{\eta_1(x)}, \bar{\eta_2(x)}\} = min\{\bar{\eta_1(x)}, 1 - \bar{\eta_1(x)}\}.$$

when $\bar{\eta_1(x)} + \bar{\eta_2(x)} = 1$;
Now, we can write the conditional risk of NN with $r^*(x)$

$$r(x) = 2 * \bar{\eta_1(x)} * \bar{\eta_2(x)} = 2 * \bar{\eta_1(x)} * (1 - \bar{\eta_1(x)}) = 2r^*(x)(1 - r^*(x))$$

till this point , we have shown in large sample case with probability 1 , a random observation x will correctly classified with probability $2r^*(x)(1-r^*(x))$.

we have shown in large sample case , that with probability one a randomly chosen x will be correctly classified with probability $2r^*(x)(1 - r^*(x))$.
For the overall NN risk R we have , by the definition

$$R = lim_n E[r(x, x_n^{'})]$$

so applying dominated convergence theorem, we ca switch order of the limit and expectation and get :

$$R = E[lim_n r(x, x_n^{'})]$$

from applying the limit ,yields

$$R = E[r(x)] = E[2 * \bar{\eta_1(x)} * \bar{\eta_2(x)}] = E[2r^*(x)(1 - r^*(x))]$$

Proof of upper-bound :

$$R = E[2r^*(x)(1 - r^*(x))] = 2E[r^*(x)] - 2E[r^*(x)^2]$$

$$Var(X) = E[X^2] - E[X]^2$$

Then we can rewrite it :

$$-2E[r^*(x)^2] = -2(Var(r^*(x)) + E[r^*(x)]^2)$$

$$R = 2E[r^*(x)] - 2(Var(r^*(x)) + E[r^*(x)]^2)$$

Depend on the definition Bayes risk we get :

$$E[r^*(x)] = R^*$$

$$R = 2R^* - 2{R^*}^2 - 2Var(r^*(x))$$

The variance $Var(r^*(x)) \geq 0$

$$R \leq 2R^*(1 - R^*)$$

Proof of lower-bound :

$$R = E[2r^*(x)(1-r^*(x))] = 2E[r^*(x)+r^*(x)(1-2r^*(x))] = R^*+E[r^*(x)(1-2r^*(x))] \geq R^*$$