ISSN 2347-1921

On the Bounds of the Expected Nearest Neighbor Distance

Mohamed M. Rizk
Mathematics & Statistics Department
Faculty of Science, Taif University, Taif, Saudi Arabia
Permanent Address: Mathematics Department,
Faculty of Science, Menoufia University, Shebin El-Kom, Egypt
mhm96@yahoo.com

Azhari A. Elhag
Mathematics & Statistics Department
Faculty of Science, Taif University, Taif, Saudi Arabia
azhri_elhag@hotmai.com

Saieed F. Ateya
Mathematics & Statistics Department
Faculty of Science, Taif University, Taif, Saudi Arabia
Permanent Address: Mathematics Department,
Faculty of Science, Assiut University, Assiut, Egypt
said_f_atya@yahoo.com

ABSTRACT

In this paper, we give some contributions for special distributions having unbounded support $S = (-\infty, \infty)$ for which we derive upper and lower bounds on the expected nearest neighbor distance of the extreme value (Gumbel) distribution as typical. Then we found the risk of nearest neighbor classifier of this distribution.

KEYWORDS

Nearest neighbor rule, expected nearest neighbor distance, extreme value distribution.

ISSN 2347-1921

1 INTRODUCTION

One of the oldest and simplest methods for pattern classification is the nearest neighbors rule; it was first studied by Fix and Hodges [6], [7]. Cover and Hart [1] proved that $R^* \leq R_\infty \leq 2R^*(1-R^*)$ under certain conditions, where R^* denotes the Bayes error, and R_∞ is the nearest neighbor risk in the infinite-sample limit. Cover [2] has shown that $R_m = R_\infty + O(m^{-2})$ for the nearest neighbor classifier in the case one-dimensional bounded support, mixture density $f \geq c > 0$, and under some additional conditions, where R_m denotes the finite sample risk, and m is the sample size. Kulkarni and Posner [10] studied the rate of convergence for nearest neighbor estimation in terms of the covering numbers of totally bounded sets. Evans et al. [5] derived an asymptotic moments of near neighbor distance distributions. Irle and Rizk [9] found an asymptotic evaluation of the conditional risk $R_m(x)$ (the probability of error conditioned on the event that $X = x$) by using partial integration and Laplace's method. Liitiäinen et al. [11] studied a boundary corrected expansion of the moments of nearest neighbor distributions. Rizk and Ateya [12] found lower and upper bounds for the risk of nearest neighbor of generalized exponential distribution.Rizk [13] found lower and upper bounds on the expected nearest neighbor distancefor exponential distribution. Rizk [14] found lower and upper bounds on the expected nearest neighbor distancefor logistic and Laplace distributions.

In this paper, we find upper and lower bounds on the expected nearest neighbor distance for distributions having unbounded support $S = (-\infty, \infty)$ for which we derive upper and lower bounds on the expected nearest neighbor distance of extreme value distribution as typical. Then we found the risk of nearest neighbor classifier of this distribution.

In pattern recognition if we have a random variable (X, θ), such that $X \in R^d$ is an observed pattern from which we wish to predict the unobservable class θ. This class takes values in a finite set $M = \{1,2,\ldots,C\}$. If the joint distribution of (X,θ) is known, then the best classifier is the Bayes classifier, see [4], [8]. In general the joint distribution of (X,θ) will be unknown, and we have a training sequence $D_m = \left((X^{(1)},\theta^{(1)}),(X^{(2)},\theta^{(2)}),\ldots,(X^{(m)},\theta^{(m)})\right)$ at our disposal, where patterns and corresponding classes are observed and we assume that $\left((X^{(1)},\theta^{(1)}),(X^{(2)},\theta^{(2)}),\ldots,(X^{(m)},\theta^{(m)})\right)$, the data, stem from a sequence of independent identically distributed random pairs with the same distribution as (X,θ). The nearest neighbor rule is an easy classification technique, classifies new observations into their appropriate categories by simply searching for similar or closest instances in the well known classified observations (training sequence). Closeness is defined in terms of a distance metric, such as Euclidean distance.

The nearest neighbor rule assigns any input feature vector to the class given by the label θ' of the nearest reference vector. The problem to be considered is the classification of a random variable θ taking values in $M = \{1,2\}$ given a sample X in χ, with the goal of minimizing the finite-sample risk $R_m = P(\theta \neq \theta')$, where χ is a separable metric space equipped with metric ρ which we denote as the pair (χ,ρ). Define the nearest distance at time m as $d_m = \rho(X, X')$.

2 BOUNDS ON THE EXPECTED NEAREST NEIGHBOR DISTANCE FOR THE EXTREME VALUE DISTRIBUTION

Let X have a probability density function $\frac{1}{\beta} e^{-\frac{x-\alpha}{\beta}} e^{-e^{-\frac{x-\alpha}{\beta}}}$, $-\infty < x < \infty$,where α the location parameter and $b > 0$ the scale parameter. Now, without loss of the generality, we assume that X have a probability density function$f(x) = e^{-x}e^{-e^{-x}}$, $-\infty < x < \infty$.

2.1 An upper bound on the expected nearest neighbor distance for the extreme value distribution

We use constants, $-\infty < K_1(m) \leq 0 \leq K_2(m) < \infty$ depending on m, to write

$$Ed_m = \int_{-\infty}^{\infty}\int_0^{\infty} P(\ |X-x|> \varepsilon)^m \, d\varepsilon \, f(x)dx$$

$$= \int_{-\infty}^{K_1(m)}\int_0^{\infty} P(|X-x| > \varepsilon)^m d\varepsilon f(x)dx$$

$$+ \int_{K_2(m)}^{\infty}\int_0^{\infty} P(\ |X-x|> \varepsilon)^m d\varepsilon f(x)dx$$

$$+ \int_{K_1(m)}^{K_2(m)}\int_0^{\infty} P(\ |X-x|> \varepsilon)^m d\varepsilon f(x)dx.$$

$$= L_1(m) + L_2(m) + L_3(m). \qquad (2.1)$$

where,

$$L_1(m) = \int_{-\infty}^{K_1(m)}\int_0^{\infty} P(|X-x| > \varepsilon)^m d\varepsilon f(x)\, dx, (2.2)$$

$$L_2(m) = \int_{K_2(m)}^{\infty}\int_0^{\infty} P(\ |X-x|> \varepsilon)^m d\varepsilon f(x)dx, (2.3)$$

ISSN 2347-1921

$$L_3(m) = \int_{K_1(m)}^{K_2(m)} \int_0^{\infty} P(\quad |X - x|> \varepsilon)^m d\varepsilon f(x)dx. (2.4)$$

Firstly, we evaluate $L_1(m)$and $L_2(m)$, and assume that$-K_1(m) = K_2(m) > 0$, for $x \in R, t > 0$.

By Markov's inequality for any $0 < t < 1$

$$\int_0^{\infty} P(\quad |X - x|> \varepsilon)^m \ d\varepsilon = \int_0^{\infty} P\left(e^{t|X-x|} > e^{t\varepsilon}\right)^m \ d\varepsilon$$

$$\leq \int_0^{\infty} \quad \varphi(t,x)^m e^{-mt\varepsilon} \ d\varepsilon = \frac{1}{mt}\varphi(t,x)^m,$$

where,$\varphi(t,x) = E\left(e^{t|X-x|}\right)$.Hence for $t = \frac{1}{\tau m}, \tau \geq 1$, we have

$\int_0^{\infty} P(|X - x| > \varepsilon)^m d\varepsilon \leq \tau\varphi\left(\frac{1}{\tau m}, x\right)^m$ .It follows

$$L_1(m) \leq \tau \int_{-\infty}^{K_1(m)} \varphi\left(\frac{1}{\tau m}, x\right)^m f(x)dx. (2.5)$$

$$L_2(m) \leq \tau \int_{K_2(m)}^{\infty} \varphi\left(\frac{1}{\tau m}, x\right)^m f(x)dx. (2.6)$$

Now, we evaluate $\varphi(t,x)$, that is we find the moment generating function of $|X - x|$. For $x \in R, 0 < t < 1$, we have

$$\varphi(x,t) = E\left(e^{t|X-x|}\right) \leq E\left(e^{tx+tX}\right) = e^{tx}E\left(e^{tX}\right) = e^{tx}\int_{-\infty}^{\infty} e^{ty} \ e^{-y}e^{-e^{-y}} \ dy$$

$$\leq e^{tx}\int_0^{\infty} u^{-t}e^{-u} \ du = e^{tx} \ \Gamma(1-t), \qquad t < 1.$$

Hence, for $t = \frac{1}{2m}$, we obtain $\varphi\left(\frac{1}{\mathrm{T}m}, x\right) \leq e^{\frac{x}{2m}}\,\Gamma\left(1 - \frac{1}{2m}\right)$. Therefore

$$\varphi\left(\frac{1}{\mathrm{T}m}, x\right)^m \leq e^{\frac{x}{2}}\left[\Gamma\left(1 - \frac{1}{2m}\right)\right]^m.$$

Now, we evaluate $L_1(m)$, from (2.2)and (2.5)we have

$$L_1(m) = \int_{-\infty}^{K_1(m)} \int_0^{\infty} P(|X - x| > \varepsilon)^m d\varepsilon\, f(x)dx$$

$$\leq 2\int_{-\infty}^{K_1(m)} \varphi\left(\frac{1}{\mathrm{T}m}, x\right)^m f(x)dx$$

$$\leq 2\left[\Gamma\left(1 - \frac{1}{2m}\right)\right]^m \int_{-\infty}^{K_1(m)} e^{\frac{x}{2}} \ e^{-x}e^{-e^{-x}}dx$$

$$\leq \frac{2}{e}\left[\Gamma\left(1 - \frac{1}{2m}\right)\right]^m \int_{-\infty}^{K_1(m)} e^{\frac{x}{2}}dx,$$

By using the inequality $\max_a a\, e^{-ma} < \frac{1}{me}$, $a > 0$, we have $e^{-x}e^{-e^{-x}} < \frac{1}{e}$, since $e^{-x} > 0$, for all x, then

$$L_1(m) \leq \frac{4}{e}\left[\Gamma\left(1 - \frac{1}{2m}\right)\right]^m e^{\frac{K_1(m)}{2}}$$

For $K_1(m) = -2\log m$, it follows

$$L_1(m) \le \frac{4}{e}\left[\Gamma\left(1-\frac{1}{2m}\right)\right]^m \frac{1}{m} = O\left(\frac{1}{m}\right) (2.7)$$

Since $\left[\Gamma\left(1-\frac{1}{2m}\right)\right]^m < 2 \quad as\, m \to \infty.$

Similarly, from(2.3)and (2.6),we can evaluate $L_2(m)$

$$L_2(m) = \int_{K_2(m)}^{\infty}\int_0^{\infty} P(|X-x|>\varepsilon)^m f(x)d\varepsilon dx$$

$$\le 2\int_{K_2(m)}^{\infty} \varphi\left(\frac{1}{\tau m}, x\right)^m f(x)dx$$

$$\le 2\left[\Gamma\left(1-\frac{1}{2m}\right)\right]^m \int_{K_2(m)}^{\infty} e^{\frac{x}{2}}\, e^{-x}e^{-e^{-x}}dx$$

$$= 2\left[\Gamma\left(1-\frac{1}{2m}\right)\right]^m \int_{K_2(m)}^{\infty} e^{-\frac{x}{2}}e^{-e^{-x}}dx.$$

Then, by partial integration

$\int_{K_2(m)}^{\infty} e^{-\frac{x}{2}}e^{-e^{-x}}dx = -2e^{-e^{-x}}e^{-\frac{x}{2}}\Big|_{K_2(m)}^{\infty} + 2\int_{K_2(m)}^{\infty} e^{-\frac{x}{2}}e^{-x}e^{-e^{-x}}dx.$

Then

$$L_2(m) \le 2\left[\Gamma\left(1-\frac{1}{2m}\right)\right]^m \left[-2e^{-e^{-x}}e^{-\frac{x}{2}}\Big|_{K_2(m)}^{\infty} + 2\int_{K_2(m)}^{\infty} e^{-\frac{x}{2}}e^{-x}e^{-e^{-x}}dx\right]$$

$$\le 2\left[\Gamma\left(1-\frac{1}{2m}\right)\right]^m \left(2e^{-e^{-K_2(m)}}e^{-\frac{K_2(m)}{2}} + \frac{2}{e}\int_{K_2(m)}^{\infty} e^{-\frac{x}{2}}\,dx\right)$$

$$= 2\left[\Gamma\left(1-\frac{1}{2m}\right)\right]^m \left(2e^{-e^{-K_2(m)}}e^{-\frac{K_2(m)}{2}} + \frac{4}{e}e^{-\frac{K_2(m)}{2}}\right).$$

For $K_2(m) = 2\log m$

$$L_2(m) \le 4\left[\Gamma\left(1-\frac{1}{2m}\right)\right]^m \left(\frac{e^{-\frac{1}{m^2}}}{m} + \frac{2}{em}\right)$$

$= \frac{4}{m}\left[\Gamma\left(1-\frac{1}{2m}\right)\right]^m \left(e^{-\frac{1}{m^2}} + 2e^{-1}\right).$ (2.8)

Now, we evaluate $L_3(m)$.From (2.4), we have

$$L_3(m) = \int_{K_1(m)}^{K_2(m)}\int_0^{\infty} P(|X-x|>\varepsilon)^m d\varepsilon\, f(x)dx$$

$$= \int_{K_1(m)}^{K_2(m)}\int_0^{\infty} e^{-mG(x,\varepsilon)}\, f(x)\, d\varepsilon\, dx, (2.9)$$

where $G(x,\varepsilon) = -\log P(|X-x|>\varepsilon)$.

Since, $-\log\ (1-y) \ge y$for all $0 \le y \le 1$,then

$-\log P(|X-x|>\varepsilon) = -\log\big(1-P(|X-x|\le\varepsilon)\big) \ge P(|X-x|\le\varepsilon)$

ISSN 2347-1921

$$= P(x - \varepsilon \le X \le x + \varepsilon) = F(x+\varepsilon) - F(x-\varepsilon)(2.10)$$

Then we need good asymptotic estimates for $F(x+\varepsilon) - F(x-\varepsilon)$, as $(\varepsilon \to 0)$, By using the Taylor expansion for the functions $F(x+\varepsilon)$ and $F(x-\varepsilon)$ we obtain

$$F(x+\varepsilon) = F(x) + \frac{f(x)\varepsilon}{1!} + \frac{f'(x)\varepsilon^2}{2!} + \frac{f''(x)\varepsilon^3}{3!} + \frac{f'''(x)\varepsilon^4}{4!} + \frac{f^{(4)}(x)\varepsilon^5}{5!} + \cdots, \qquad (2.11)$$

$$F(x-\varepsilon) = F(x) - \frac{f(x)\varepsilon}{1!} + \frac{f'(x)\varepsilon^2}{2!} - \frac{f''(x)\varepsilon^3}{3!} + \frac{f'''(x)\varepsilon^4}{4!} - \frac{f^{(4)}(x)\varepsilon^5}{5!} + \cdots. \quad (2.12)$$

Substituting (2.11) and (2.12) in (2.10) yields

$$F(x+\varepsilon) - F(x-\varepsilon) = \frac{2f(x)\varepsilon}{1!} + \frac{2f''(x)\varepsilon^3}{3!} + \frac{2f^{(4)}(x)\varepsilon^5}{5!} + \cdots \ge 2\varepsilon\, f(x),$$

since$f^{(n)}(x) \ge 0$ for $n = 0, 2, 4, \ldots$, then we obtain $G(x,\varepsilon) \ge 2\varepsilon\, f(x)$. Hence

$$L_3(m) \le \int_{K_1(m)}^{K_2(m)} \int_0^{\infty} e^{-2m\varepsilon f(x)} f(x)\, d\varepsilon dx$$

$$= \int_{K_1(m)}^{K_2(m)} \frac{1}{2m} dx = \frac{1}{2m}\big(K_2(m) - K_1(m)\big).$$

$$= \frac{1}{2m}(\log m^2 + \log m^2) \le \frac{2\log m}{m}. (2.13)$$

Substituting (2.8), (2.9) and (2.13) in (2.1), we obtain the upper bound of Ed_m for the extreme value distribution

$$Ed_m \le \frac{4}{e}\left[\Gamma\left(1 - \frac{1}{2m}\right)\right]^m \frac{1}{m} + \left[\Gamma\left(1 - \frac{1}{2m}\right)\right]^m \frac{4}{m}\left(e^{-\frac{1}{m^2}} + 2e^{-1}\right) + \frac{2\log m}{m}$$

$$\le \frac{4}{m}\left[\Gamma\left(1 - \frac{1}{2m}\right)\right]^m \left(e^{-\frac{1}{m^2}} + 3e^{-1}\right) + \frac{2\log m}{m}. (2.14)$$

2.2 A lower bound on the expected nearest neighbor distance for the extreme value distribution

In this section we derive the lower bounds for expected nearest neighbor distance Ed_m for the extreme value distribution with the density $f(x) = e^{-x}e^{-e^{-x}}, -\infty < x < \infty$.Then

$$Ed_m = \int_0^{\infty} P(d_m > \varepsilon)\, d\varepsilon = \int_{-\infty}^{\infty}\int_0^{\infty} P(d_m > \varepsilon | X = x)\, d\varepsilon\, f(x) dx$$

$$= \int_{-\infty}^{\infty}\int_0^{\infty} P(\ |X - x| > \varepsilon\ )^m\ d\varepsilon\, f(x)dx$$

$$= \int_{-\infty}^{\infty}\int_0^{\infty} P(\ |X - x| > \varepsilon\ )^m\, d\varepsilon\, e^{-x}e^{-e^{-x}} dx$$

$$\ge \int_{-\infty}^{\infty}\int_0^{\infty} P(X < x - \varepsilon)^m\, d\varepsilon\ \ e^{-x}e^{-e^{-x}} dx$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{x} P(X < z)^m\, dz\ \ e^{-x}e^{-e^{-x}} dx$$

$$= \int_{-\infty}^{\infty}\int_{z}^{\infty} e^{-x}e^{-e^{-x}}\, dx\, P(X < z)^m dz$$

$$= \int_{-\infty}^{\infty} \left(1 - e^{-e^{-z}}\right)\left(e^{-me^{-z}}\right) dz$$

$$= \int_0^{\infty} \left(e^{-mt} - e^{-(m+1)t}\right) t^{-1}\, dt$$

$$\ge \int_0^{\infty} \left(e^{-mt} - e^{-(m+1)t}\right) e^{-t}\, dt$$

$$= \frac{1}{(m+1)(m+2)}. (2.15)$$

Note that, from the distribution has exponentially decaying tails there is an additional logarithmic term over the rates for compact support. This example illustrates that the expected nearest neighbor distance depends on the tails of the distribution.

Now, we can find an upper bound on the finite sample risk $R_m$in terms of the expected nearest neighbor distance for extreme value distribution by using the following result:

If, for some $\omega_1 > 0$ and$0 < \gamma \leq 1$ we have $|m(x) - m(x')| \leq \omega_1 \rho(x, x')^\gamma$, for all $x, x' \in \chi$, then for some suitable $\omega > 0$independent of m,

$$R_m \leq R_\infty + \omega\left[(Ed_m)^\gamma + \left(Ed_m^{2\gamma}\right)\right], (2.16)$$

where $\omega = max\{\omega_1, \omega_1^2\}$.

This result is due to Irle and Rizk [9], for which they found an upper bound on the finite sample risk $R_m$in terms of the expected nearest neighbor distance.

Putting $\gamma = \frac{1}{2}$in (2.16), we obtain $R_m \leq R_\infty + \omega[\sqrt{Ed_m} + Ed_m]$.

Hence, from (2.13) we have

$$R_m \leq R_\infty + \lambda\sqrt{\frac{4}{m}\left[\Gamma\left(1 - \frac{1}{2m}\right)\right]^m \left(e^{-\frac{1}{m^2}} + 2e^{-1}\right) + \frac{2\log m}{m}} + \lambda\left[\frac{4}{m}\left[\Gamma\left(1 - \frac{1}{2m}\right)\right]^m \left(e^{-\frac{1}{m^2}} + 3e^{-1}\right) + \frac{2\log m}{m}\right]$$

$$\leq R_\infty + \lambda\left[\sqrt{\frac{C_2}{m} + \frac{4\log m}{2m}} + \frac{C_2}{m} + \frac{2\log m}{m}\right], (2.17)$$

where$C_2 = 4\left[\Gamma\left(1 - \frac{1}{2m}\right)\right]^m \left(e^{-\frac{1}{m^2}} + 2e^{-1}\right)$.Note that C_2 dependent on m.

REFERENCES

[1] Cover, T. M. and Hart, P. E. 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13: 21-27.

[2] Cover, T. M. 1968.Rates of convergence for nearest neighbor procedures. In Proceedings of the Hawaii International Conference on Systems Sciences, pages 413-415. Honolulu, HI.

[3] Dasarathy, B. 1991. Nearest neighbor classification techniques.IEEE, Los Alamitos, CA.

[4] Devroye, L., Györfi, L., and Lugosi, G. 1996. Probabilistic theory of pattern recognition. Springer-Verlag, New York.

[5] Evans, D., Jones, A. J., and Schmidt, W. M. (2002). Asymptotic moments of near neighbor distance distributions. Proc. R. Soc. Lond. A 458, 2839-2849.

[6] Fix, E., Hodges, J. 1991. Discriminatory analysis, nonparametric discrimination, consistency properties.in nearest neighbor classification techniques, Dasarathy, B., editor, pages 32-39, IEEE Computer Society Press, Los Alamitos, CA.

[7] Fix, E., Hodges, J. 1991. Discriminatory analysis, small sample performance.in nearest neighbor classification techniques, Dasarathy, B., editor, pages 40-56, IEEE Computer Society Press, Los Alamitos, CA.

[8] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. 2002. A distribution-free theory of nonparametric regression, Springer-Verlag, New York.

[9] Irle, A. Rizk, M. 2004. On the risk of nearest neighbor rules.DsserationzurErlangung des Doktorgrades.

[10] Kulkarni, S. R. and Posner, S. E. 1995. Rates of convergence of nearest neighbor estimation under arbitrary sampling. IEEE Transactions on Information Theory, 41: 1028-1039.

[11] Liitiäinen, E., Lendasse, A., and Corona, F. (2010). A boundary corrected expdansion of the moments of nearest neighbor distributions. Random Structures and Algorithms 37, 223-247.

[12] Rizk, M. M. and Ateya, S. F. (2013). On lower and upper bounds for the risk in pattern recognition**.**Applied Mathematical Sciences.Vol. 7, 6417-6428.

[13] Rizk, M. M. (2013). Bounds on the finite-sample risk for exponential distribution.Journal of Advances in Mathematics, Vol. 4(2), 436-440.

[14] Rizk, M. M. (2015). On the bounds of the expected nearest neighbor distance**.**British Journal of Mathematics & Computer Science, Vol. 6(4), 297-306.