ISSN 2347-1921

Bounds on the Finite-Sample Risk for Exponential Distribution

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

ABSTRACT

In this paper, we derive lower and upper bounds on the expected nearest neighbor distance for exponential distribution. Then we find bounds on the risk of the nearest neighbor for exponential distribution.

Keywords

Nearest neighbor classification; expected nearest neighbor distance; exponential distribution.

Council for Innovative Research
Peer Review Research Publishing System
Journal: Journal of Advances in Mathematics
Vol 4, No 2
editor@cirworld.com
www.cirworld.com, member.cirworld.com

1 Introduction and Model

The neatest neighbor rule is one of the simplest types of nonparametric methods of interest in statistical pattern recognition that can be used with arbitrary distributions and without the assumption that the forms of the underling densities are known.

The nearest neighbor rule was first studied by Fix and Hodges [5], [6]. Cover and Hart [1] gave upper bounds for the limit of the risk of nearest neighbor classifiers under certain conditions. 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, R_∞ is the nearest neighbor risk in the infinite-sample limit, and m is the sample size. Wagner [14] and Fritz [7] treated convergence of the conditional error rate for nearest neighbor. Fukunaga and Hummels [8] studied the rate of convergence of the above bias in d- dimensional feature space. Psaltis et al. [12] generalised the results of Cover [2] to general dimension, and Snapp and Venkatesh [13] further extended the results to the case of multiple classes. Kulkarni and Posner [11] studied the rate of convergence for nearest neighbor estimation in terms of the covering numbers of totally bounded sets. Irle and Rizk [10] 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. There is a wealth of consistency results in different directions available for nearest neighbor rules; see the collection of Dasarathy [3], the monographs by Devroye et al. [4], and Györfi et al. [9].

In this paper, we find lower and upper bounds on the expected nearest neighbor distance for exponential distribution as typical for distributions having unbounded support, and derive the bounds on the risk of nearest neighbor for a two-class pattern recognition of this distribution.

We will consider (X, θ) be a random pair taking values in $\chi \times \{1,2\}$, where X taking values in some general separable metric space χ equipped with metric ρ which we denote as the pair (χ, ρ), and let $D_m = \Big((X^{(1)}, \theta^{(1)}), (X^{(2)}, \theta^{(2)}), \dots, (X^{(m)}, \theta^{(m)}) \Big)$ be a sequence of independent identically distributed random pairs with the same distribution as (X, θ). The $X^{(i)}$ are called the observations and $\theta^{(i)}$ are usually called the classes. The function $\delta : \chi \to \{1,2\}$, where $\delta(x)$ represents one's guess of θ given x is called a classifier. The probability of error for a classifier δ is $P(\theta \neq \delta(X))$.

If the joint distribution of (X, θ) is known then the best classifier is known as the Bayes classifier. The Bayes classifier δ^* minimizes this risk resulting in the conditional Bayes risk

$$r^*(x) = P(\theta \neq \delta^*(x) | X = x) \leq P(\theta \neq \delta(x) | X = x), \quad \text{for all classifier } \delta.$$

The Bayes risk is given by

$$R^* = E\big(r^*(x)\big) = \int r^*(x) P^X dx.$$

Define the conditional mean of θ given $X = x$ as

$$m(x) = P(\theta = 1 | X = x) = E(\theta | X = x),$$

and the conditional variance as

$$\sigma^2(x) = P(\theta = 1 | X = x) - [P(\theta = 1 | X = x)]^2.$$

In general the joint distribution of (X, θ) is unknown it is often assumed that in addition to X we have a training sequence $D_m = \Big((X^{(1)}, \theta^{(1)}), (X^{(2)}, \theta^{(2)}), \dots, (X^{(m)}, \theta^{(m)}) \Big)$, where patterns corresponding classes observed and we assume that $\Big((X^{(1)}, \theta^{(1)}), (X^{(2)}, \theta^{(2)}), \dots, (X^{(m)}, \theta^{(m)}) \Big)$, the data, stem from a sequence of independent identically distributed random pairs with the same distribution as (X, θ).

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 $\{1,2\}$ given a sample X in χ, with the goal of minimizing the finite-sample risk $R_m = P(\theta \neq \theta')$. The conditional probability of error for the nearest neighbor rule is defined as the probability of error in classification θ by θ' given X and its nearest neighbor X' and denoted by $P(\theta \neq \theta' | X, X')$. By averaging $P(\theta \neq \theta' | X, X')$ over X', we obtain the m-samples conditional average probability of error $R_m(X) = P(\theta \neq \theta' | X)$, and by averaging $P(\theta \neq \theta' | X)$ with respect to X, we obtain the unconditional nearest neighbor risk (the unconditional probability of error)

$$R_m = P(\theta \neq \theta') = \int P(\theta \neq \theta' | X) \, f(x) dx.$$

Define the nearest distance at time m as $d_m = \rho(X, X')$.

In the next, we begin by presenting (without proof) the following result, this result is due to Irle and Rizk [10], for which they found an upper bound on the finite sample risk R_m in terms of the expected nearest neighbor distance.

Lemma 1.

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

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

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

2. A Bound Risk of the Expected Nearest Neighbour Distance for Exponential Distribution

In this section we derive the lower and upper bounds for the expected nearest neighbor distance Ed_m for exponential distribution. Let X has a density function $f(x) = \lambda e^{-\lambda x}$, $\lambda, x > 0$.

2.1 Deriving a lower bound

$$\begin{aligned} Ed_m &= \int_0^\infty P(d_m > \varepsilon)\, d\varepsilon = \int_0^\infty \int_0^\infty P(d_m > \varepsilon | X = x)\, d\varepsilon\, f(x) dx \\ &= \int_0^\infty \int_0^\infty P(\, |X - x| > \varepsilon\,)^m \, d\varepsilon\, f(x) dx \\ &= \int_0^\infty \int_0^\infty P(\, |X - x| > \varepsilon\,)^m \, d\varepsilon\, \lambda e^{-\lambda x} dx \qquad (2.1.1) \\ &\ge \int_0^\infty \int_0^\infty P(X < x - \varepsilon)^m \, d\varepsilon\, \lambda e^{-\lambda x} dx \\ &= \int_0^\infty \int_0^x P(X < z)^m \, dz\, \lambda e^{-\lambda x} dx \\ &= \int_0^\infty \int_0^x \left(1 - e^{-\lambda z}\right)^m dz\, \lambda e^{-\lambda x} dx \\ &= \int_0^\infty \int_z^\infty \lambda e^{-\lambda x} dx\, \left(1 - e^{-\lambda z}\right)^m dz \\ &= \int_0^\infty e^{-\lambda z} \left(1 - e^{-\lambda z}\right)^m dz = \frac{1}{\lambda} \int_0^\infty \left(1 - e^{-\lambda z}\right)^m d\left(1 - e^{-\lambda z}\right) dz \\ &= \frac{1}{\lambda} \int_0^1 y^m\, dy = \frac{1}{\lambda(m+1)}. \qquad (2.1.2) \end{aligned}$$

2.2 Deriving an upper bound:

We use a constant $0 \le K_1(m) < \infty$ depending on m, to write

$$\begin{aligned} Ed_m &= \int_0^\infty \int_0^\infty P(\, |X - x| > \varepsilon\,)^m \, d\varepsilon\, f(x) dx \\ &= \int_0^{K_1(m)} \int_0^\infty P(|X - x| > \varepsilon)^m \, d\varepsilon\, f(x) dx + \int_{K_1(m)}^\infty \int_0^\infty P(|X - x| > \varepsilon)^m \, d\varepsilon\, f(x) dx \\ &= L_1(m) + L_2(m), \qquad (2.2.1) \end{aligned}$$

where

$$L_1(m) = \int_0^{K_1(m)} \int_0^\infty P(|X - x| > \varepsilon)^m \, d\varepsilon\, f(x) dx, \qquad (2.2.2)$$

$$L_2(m) = \int_{K_1(m)}^\infty \int_0^\infty P(\, |X - x| > \varepsilon\,)^m \, d\varepsilon f(x) dx. \qquad (2.2.3)$$

Firstly, we evaluate $L_1(m)$.

We write, for X with density f

$$\begin{aligned} L_1(m) &= \int_0^{K_1(m)} \int_0^\infty P(|X - x| > \varepsilon)^m d\varepsilon\, f(x) dx \\ &= \int_0^{K_1(m)} \int_0^\infty e^{-mG(x,\varepsilon)} f(x) d\varepsilon\, dx, \end{aligned}$$

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

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

$$-\log P(|X - x| > \varepsilon) = -\log\left(1 - P(|X - x| \le \varepsilon)\right)$$

$$\geq P(|X-x| \leq \varepsilon) = P(x-\varepsilon \leq X \leq x+\varepsilon)$$
$$= F(x+\varepsilon) - F(x-\varepsilon). \qquad (2.2.4)$$

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

$$F(x+\varepsilon) = F(x) + \frac{f(x)\varepsilon}{1!} + \frac{f'(x)\varepsilon^2}{2!} + \frac{f''(x)\varepsilon^3}{3!} + \frac{f'''(x)\varepsilon^4}{4!} + \frac{f^{(4)}(x)\varepsilon^5}{5!} + \cdots, \qquad (2.2.5)$$

$$F(x-\varepsilon) = F(x) - \frac{f(x)\varepsilon}{1!} + \frac{f'(x)\varepsilon^2}{2!} - \frac{f''(x)\varepsilon^3}{3!} + \frac{f'''(x)\varepsilon^4}{4!} - \frac{f^{(4)}(x)\varepsilon^5}{5!} + \cdots. \qquad (2.2.6)$$

Substituting $(2.2.5)$ and $(2.2.6)$ in $(2.2.4)$ yields

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

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

$$L_1(m) \leq \int_0^{K_1(m)} \int_0^{\infty} e^{-2m\varepsilon f(x)} f(x) d\varepsilon\, dx = \int_0^{K_1(m)} \frac{1}{2m}\, dx = \frac{K_1(m)}{2m} \qquad (2.2.7)$$

Now, we evaluate $L_2(m)$. By Markov's inequality for any $0 < t < 1$

$$\int_0^{\infty} P(\,|X-x| > \varepsilon)^m\, d\varepsilon = \int_0^{\infty} P\left(\, e^{t|X-x|} > e^{t\varepsilon}\right)^m d\varepsilon$$
$$\leq \int_0^{\infty} \varphi(t,x)^m e^{-mt\varepsilon}\, d\varepsilon = \frac{1}{mt}\, \varphi(t,x)^m,$$

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

$$L_2(m) \leq \tau \int_{K_1(m)}^{\infty} \varphi\left(\frac{1}{\tau m}, x\right)^m f(x)dx. \qquad (2.2.8)$$

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

$$\varphi(t,x) = E\left(e^{t|X-x|}\right) \leq e^{tx} E(e^{tX}) = e^{tx} \int_0^{\infty} e^{ty}\, \lambda e^{-\lambda y}\, dy$$
$$= \lambda e^{tx} \int_0^{\infty} e^{-y(\lambda-t)} dy = \frac{\lambda e^{tx}}{\lambda - t},\ t < \lambda$$

Hence for $t = \frac{1}{\tau m}$, $\tau \geq 1$

$\varphi\left(\frac{1}{\tau m}, x\right)^m \leq e^{\frac{x}{\tau}} \left(\frac{\lambda}{\lambda - t}\right)^m = e^{\frac{x}{\tau}} \left(\frac{1}{1 - \frac{1}{\tau m\lambda}}\right)^m = e^{\frac{x}{\tau}} \left(1 + \frac{1}{\tau m\lambda - 1}\right)^m$. It follows

$$L_2(m) \leq \tau\lambda \left(1 + \frac{1}{\tau m\lambda - 1}\right)^m \int_{K_1(m)}^{\infty} e^{\frac{x}{\tau}}\, e^{-\lambda x}\, dx$$
$$= \frac{\tau^2\lambda}{\tau\lambda - 1} \left(1 + \frac{1}{\tau m\lambda - 1}\right)^m e^{-(\tau\lambda - 1)K_1(m)},\ \ \tau\lambda > 1. \qquad (2.2.9)$$

Substituting $K_1(m) = \frac{log\ m}{(\tau\lambda - 1)}$ in $(2.2.7)$ and $(2.2.9)$, we obtain

$$L_1(m) \leq \frac{log\ m}{2(\tau\lambda - 1)\, m}, \qquad (2.2.10)$$

$$L_2(m) \leq \frac{\tau^2\lambda}{\tau\lambda - 1} \left(1 + \frac{1}{\tau m\lambda - 1}\right)^m \frac{1}{m} = O\left(\frac{1}{m}\right), \qquad (2.2.11)$$

since $\left(1 + \frac{1}{\tau m\lambda - 1}\right)^m \to e^{\frac{1}{\tau\lambda}},\ as\ m \to \infty.$

Substituting $(2.2.10)$ and $(2.2.11)$ in $(2.2.1)$ yields

$$Ed_m = L_1(m) + L_2(m) \leq O\left(\frac{1}{m}\right) + \frac{log\ m}{2(\tau\lambda - 1)\ m}\ , \tau\lambda > 1. \qquad (2.2.12)$$

Note that, from $(2.1.2)$ and $(2.2.12)$ the lower and upper bounds of the expected nearest neighbor distance are different in constants and the term $log\ m.$ That is, for the distributions have 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.

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

Hence, from $(2.2.12)$ we have

$$R_m \leq R_\infty + \omega\left[\left(O\left(\frac{1}{m}\right) + \frac{log\ m}{c_2(\tau\lambda - 1)m}\right)^{\frac{1}{2}} + \left(O\left(\frac{1}{m}\right) + \frac{log\ m}{c_2(\tau\lambda - 1)m}\right)\right]. \qquad (2.2.13)$$

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] 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.

[6] 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.

[7] Fritz, J. 1975. Distribution-free exponential error bound for nearest neighbor pattern classi_cation. IEEE Trans. Inform. Theory 21, 552-557.

[8] Fukunaga, K. and Hummels, D. M. 1987. Bias of nearest neighbor estimates. IEEE Trans. Pattern Anal. Mach. Intell., 9: 103-112.

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

[10] Irle, A. Rizk, M. 2004. On the risk of nearest neighbor rules. Dsseration zur Erlangung des Doktorgrades.

[11] 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.

[12] Psaltis, D., Snapp, R, R. and Venkatesh, S. S. 1994. On the finite sample performance of the nearest neighbor classifier. IEEE Transaction on Information Theory, 40: 820-837.

[13] Snapp, R. R. and Venkatesh, S. S. 1998. Asymptotic expansion of the K-nearest neighbor risk. Ann. Statist. 26: 850-878.

[14] WAGNER, T.J. 1971. Convergence of the nearest neighbor rule. IEEE Trans. Inform. Theory 17, 566-571.