Bayesian Diffusion Decision Model for Adaptive k-Nearest Neighbor Classification

Yung-Kyun Noh[⋆†], Frank Chongwoo Park[†], Daniel D. Lee[⋆]

[⋆]GRASP Lab., Department of Electrical and Systems Engineering,
University of Pennsylvania, Philadelphia, PA 19104

[†]Robotics Lab. School of Mechanical and Aerospace Engineering
Seoul National University, Seoul 151-746, Korea

nohyung@seas.upenn.edu, fcp@snu.ac.kr, ddlee@seas.upenn.edu

Topic: Learning algorithms/biological learning, Preference: Oral/Poster

Nearest neighbor classification is a well-known algorithm with theoretical bounds on the classification error in the asymptotic limit (Cover and Hart, 1967). For *k-NN* classification, as the number k of nearest neighbors approaches infinity, a simple majority voting rule can achieve the optimal Bayes error in performance (Devroye et al., 1996). Unfortunately, due to sampling issues and computational complexity, using a large number k of nearest neighbors may not be ideal. Thus, previous work has proposed methods to adaptively determine the number of nearest neighbors to use for classification (Wang et al., 2006). Here we present a new Bayesian method for adaptively choosing the optimal number of nearest neighbors, based upon an exponential prior for local densities and Poisson statistics for nearest neighbor distances. We also show how this adaptive strategy can be interpreted as an optimal decision making process, closely related to diffusion decision making in psychological and neurall decision processes (Ratcliff and Mckoon, 2008).

We consider the following probabilistic model for nearest neighbors: given a local density λ, the volume u of the largest sphere inside the k-th nearest neighbor point is described by an Erlang distribution (Poczos and Schneider, 2011):

$$p(u|\lambda) = \frac{\lambda^k}{\Gamma(k)} \exp(-\lambda N u)(Nu)^{k-1} \tag{1}$$

where N is the total number of data points.

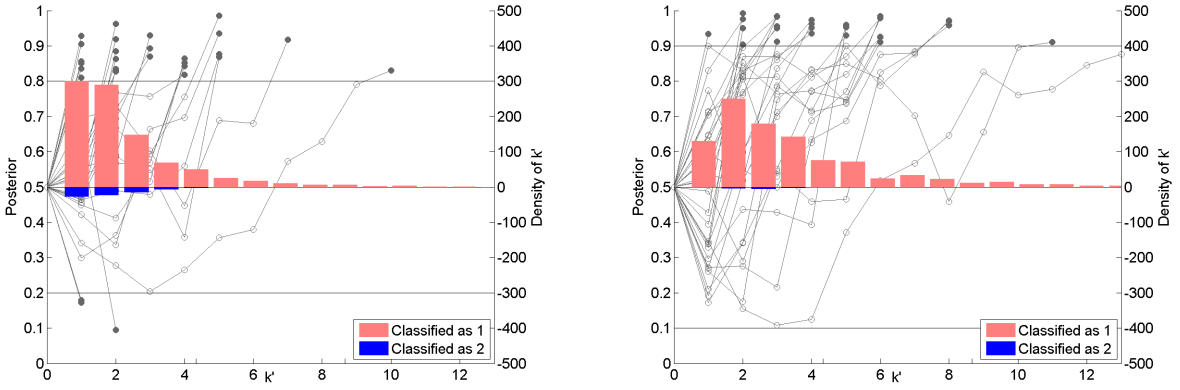
Here we describe how this distribution can be used for adaptive *k-NN* classification for two classes, with obvious extensions to multiple classes. Given the k_1-th nearest neighbor with spherical volume u_1 to class 1, and the k_2-th nearest neighbor with volume u_2 to class 2, we formulate the posterior distributions over the local class densities λ_1 and λ_2 via Bayes rule. For this analysis, we incorporate an expontial prior over local densities: $p(\lambda) = b\exp(-b\lambda)$. The resulting posteriors can be integrated to yield the posterior likelihood that λ_1 is larger than λ_2:

$$p(\lambda_1 > \lambda_2|u_1, u_2) = \tag{2}$$
$$\frac{1}{(u_1+u_2+2b)^{k_1+k_2+1}} \sum_{m=0}^{k_1} \left(\begin{array}{c} k_1+k_2+1 \\ m \end{array} \right) (u_1+b)^m (u_2+b)^{k_1+k_2+1-m}$$

It is interesting to note that this posterior is equivalent to the probability of flipping a biased coin k_1+k_2+1 times, and observing less than k_1 number of heads. The Bayesian posterior in Eq. (2) can be efficiently computed in an incremental fashion, and the nearest neighbor computation can be adaptively stopped when there is enough evidence to make a classification. Traditional *k-NN* classification is equivalent to using $k = 2k_i - 1$ as $b \to \infty$.

In Fig. 1, we show the results of a simulation where the optimal Bayes decision is $\lambda_1 > \lambda_2$. The figure shows how different realizations result in varying decision making times as the posterior likelihood becomes certain enough to shortcut the nearest neighbor computation. The incremental computation of the estimated posterior is displayed as a diffusion decision variable, which shows how some samples are immediately classified and others delay the decision until enough nearest neighbor evidence has accumulated to cross the decision threshold.

Thus, our method can be interpreted as a diffusion decision model, which has been shown to be optimal in systems with finite resources (Bogacz et al., 2006). By modulating the confidence level in the model, we can tune



(a) (b)

Figure 1: Bayesian nearest neighbor decision making process with (a) 80% confidence and (b) 90% confidence. Sample data were generated from local densities $\lambda_1 = 0.8$ and $\lambda_2 = 0.2$. The incremental computation of the posterior $P(\lambda_1 > \lambda_2)$ is displayed for different realizations, and shows the threshold where enough evidence has accumulated to make a classification. The bars represent the number of points that are correctly and incorrectly classified at each stage of the computation. Using a larger confidence results in less error, but with an concomitant increase in computation time.

the balance between using less computational resources with classification accuracy. We show how earlier models can be viewed as special cases of our general method (Beck et al., 2008; Ma et al., 2006). We also show how our methods perform on standard machine learning datasets in comparison to traditional nearest neighbor approaches.

References

A. F. Atiya. Estimating the posterior probabilities using the k-nearest neighbor rule. *Neural Computation*, 17:731–740, March 2005.

J. M. Beck, W. J. Ma, R. Kiani, T. Hanks, A. K. Churchland, J. Roitman, M. N. Shadlen, P. E. Latham, and A. Pouget. Probabilistic Population Codes for Bayesian Decision Making. *Neuron*, 60(6):1142–1152, 2008.

R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113(4):700–765, 2006.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Applications of mathematics. Springer, 1996.

C. C. Holmes and N. M. Adams. A probabilistic nearest neighbour method for statistical pattern recognition. *Journal Of The Royal Statistical Society Series B*, 64(2):295–306, 2002.

W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438, 2006.

B. Poczos and J. Schneider. On the estimation of alpha-divergences. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2011.

R. Ratcliff and G. Mckoon. The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, 20(4):873–922, 2008.

J. Wang, P. Neskovic, and L. N. Cooper. Neighborhood size selection in the k-nearest-neighbor rule using statistical confidence. *Pattern Recognition*, 39:417–423, March 2006.