Methodology and Computing in Applied Probability (2019) 21:1251–1258
https://doi.org/10.1007/s11009-018-9670-z

An Efficient Algorithm for Bayesian Nearest Neighbours

Giuseppe Nuti[1]

Received: 21 August 2017 /
Accepted: 3 September 2018 / Published online: 27 September 2018

Abstract
K-Nearest Neighbours (k-NN) is a popular classification and regression algorithm, yet one of its main limitations is the difficulty in choosing the number of neighbours. We present a Bayesian algorithm to compute the posterior probability distribution for k given a target point within a data-set, efficiently and without the use of Markov Chain Monte Carlo (MCMC) methods or simulation—alongside an exact solution for distributions within the exponential family. The central idea is that data points around our target are generated by the same probability distribution, extending outwards over the appropriate, though unknown, number of neighbours. Once the data is projected onto a distance metric of choice, we can transform the choice of k into a change-point detection problem, for which there is an efficient solution: we recursively compute the probability of the last change-point as we move towards our target, and thus *de facto* compute the posterior probability distribution over k. Applying this approach to both a classification and a regression UCI data-sets, we compare favourably and, most importantly, by removing the need for simulation, we are able to compute the posterior probability of k exactly and rapidly. As an example, the computational time for the Ripley data-set is a few milliseconds compared to a few hours when using a MCMC approach.

Keywords K-nearest neighbour · Non-parametric classification · Bayesian classification

Mathematics Subject Classification (2010) 62F15 Bayesian Inference · 60G25 Prediction Theory

1 Introduction & Related Work

Various authors have explored the idea of Bayesian k-NN algorithms, e.g. Cucala et al. (2008), Guo and Chakraborty (2010), and originally (Holmes and Adams 2002). The simplicity and elegance of k-NN lends itself, at least intuitively, to a Bayesian setting where the aim is to allow the number of neighbours to vary depending on the data (as highlighted in Ghosh (2006)).

Primarily at UBS Securities LLC, 1285 Ave. of the Americas, New York, NY 10019

✉ Giuseppe Nuti
ucacnut@ucl.ac.uk

[1] Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, UK

Practically all of the work has relied on Markov Chain Monte-Carlo methods in some form or other; the use of simulation circumvents the need to model the full joint probability distribution of the data and the number of neighbours for any target. In an attempt to avoid the use of simulation, the authors in Ji and Friel (2013) have approximated the likelihood function, albeit at the expense of accuracy and portability to regression problems. More recently, as an alternative approach, the idea of hubness is explored in Tomasev et al. (2011).

In a somewhat distinct branch of Bayesian statistics, numerous studies have focused on estimating change-point probabilities for data where the generating process is presumed to vary over time. Initial works were focused on partition analysis for the entire data-set, often using MCMC simulation: Smith (1975), Stephens (1994), Green (1995) (which, interestingly, is loosely connected with the computational complexity of estimating the posterior probability of k using approaches based on simulation). Yet it was not until the authors in Prescott Adams and MacKay (2007) presented an on-line version of Bayesian change-point estimation (with an $O(n)$ computational complexity), that change-point problems become easily estimated.

As discussed in detail by Manocha and Girolami (2007), a probabilistic view of k in nearest neighbour algorithms outperforms standard cross-validation approaches, though practitioners often avoid the probabilistic approach due to its reliance on MCMC methods of estimation. By using the algorithm presented in Prescott Adams and MacKay (2007) applied to the data ordered by distance to our target coordinates, we can compute the exact probability distribution for k specific to our target point.

2 Efficient Bayesian Nearest Neighbour

The idea of calibrating the number of neighbours to the data is centered around the notion that, within the appropriate neighbourhood, data points are similar, or, in other words, they are generated by the same process. It is indeed our goal to determine how many k neighbours represent such *appropriate* neighbourhood.

2.1 Data Generating Process

If we order the data using a distance measure of choice with respect to a target point, we have transformed our assumption into the idea that the data-generating process is shared for the first k points closets to the target. As such, moving from the most distant point towards our target, the underlying process generating the data can vary with a known probability and, when a change occurs, such process is drawn from a known prior distribution. We aim to determine the probability of such change-point having occurred at the various intervals between neighbours — once we have reached our target datum.

This formulation is equivalent to the change-point analysis in Prescott Adams and MacKay (2007): we can recursively keep track of the historical change-point probability until we reach the target point. The resulting probability of a change-point having occurred at k points from our target is indeed the probability of k being the correct number of neighbours.

2.1.1 A Simple Example

As a simple two dimensional classification problem, imagine we have the data depicted in Fig. 1, with the respective order, based on the Euclidean distance, shown below in Fig. 2.

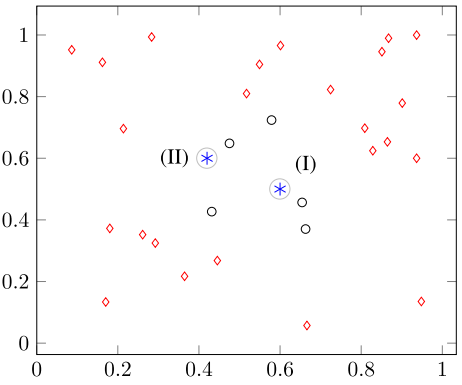

Fig. 1 Two-class data example; $k = 5$ is most probably the correct chocie for target point (I) and any choice of $k \geq 10$ will likely result in a misclassification, whilst the picture is less obvious for target point (II)

In this case, the appropriate number of neighbours for target point (I) is five. An alternative way to view the choice of k is to order the data-points by their distance to the target point (Fig. 2, below the x-axis). We note the prior probability of each k (before seeing any of the data) as the dotted line, which is just a geometric distribution with $p_\gamma = 0.05$. To compute the posterior probability for k (solid blue line in Fig. 2), we can start from the rightmost point and move towards our target: i.e. the probability that the data-generating process has changed (assuming a Beta prior probability for the data generation with parameters $\mathcal{B}(\alpha = 10., \beta = 10)$. and a probability of a change-point occurring in between any two neighbours of $p_\gamma = 0.05$, i.e. our prior on the number of neighbours is $k = 20$). Conversely, we are not as convinced of the appropriate k for target (II), as we can see form the posterior the distribution which is giving a rather mixed view.

Note that the choice of Beta distribution is the standard conjugate prior for the parameters of a binary random variable. The choice of $p_\gamma = 0.05$ is a key part of this approach. Specific to this example, we are implicitly assuming that the data will vary as we move away from our target point with a probability of 0.05. In other words, the hazard function is memoryless — with our prior for the expected number of neighbours set to $\frac{1}{p_\gamma} = 20$.

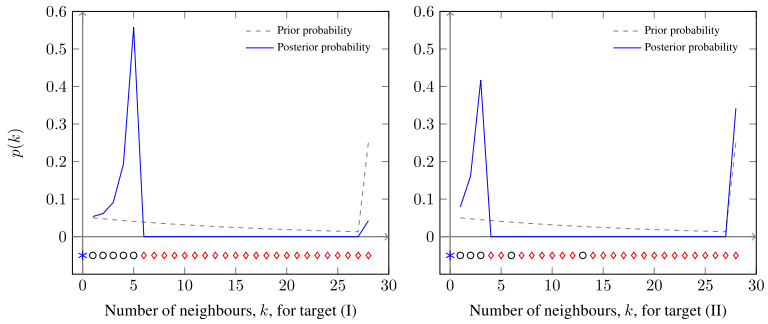

Fig. 2 Probability distribution for nearest neighbour count (with the data ordered by Euclidean distance below x-axis). The prior probability is updated into the posterior probability by observing how the data will impact our choice of k neighbours for the target (I) on the left and the target (II) on the right

2.2 Algorithm

The first step is to represent the data into an ordered list driven by the distance from our target point[1]. If we define the target point as x_τ, we order all of the available training data as $x_0, ..., x_{\tau-1}$ (defined as $\vec{x}_{0:\tau-1}$) with x_0 being the most distant point from our target. We assume that the data x_t is i.i.d. over a partition ρ from some probability distribution $P(x_t|\eta_\rho)$, where η_ρ represents the parameters of the data-generating distribution. Finally, we assume that for all of the partitions, η_ρ is also i.i.d from a known prior distribution (where $\rho = 1, ..., n$ and $n \leq \tau$). Now we are ready to use the algorithm presented in Prescott Adams and MacKay (2007), applied to the projected data[2].

Our objective is to compute the probability of each number of neighbours once we have reached our target point: $p(k_\tau{=}i|\vec{x}_{0,...,\tau-1}) \quad \forall i = 0, ..., \tau - 1$ with $\tau - 1$ total data points. Note that the subscript τ in k_τ indicates that we are representing the appropriate number of neighbours from the viewpoint of x_τ, i.e. the target point. Starting from the point farthest away, and initializing the probability of a change-point having occurred before the initial point to 1.0, we set the initial conditions[3]:

$$p(k_0 = 0) \;=\; 1.0 \tag{1}$$

$$\eta_0 \;=\; \eta_{prior} \tag{2}$$

Firstly, we note that, as we observe a new datum, moving closer to our target, the number of neighbours k_t within the same partition can either increase by one, with probability $1 - p_\gamma$, or terminate in favour of a nascent partition.

$$p(k_t|k_{t-1}) = \begin{cases} p_\gamma & \text{if } k_t = 0 \\ 1 - p_\gamma & \text{if } k_t = k_{t-1} + 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

A key advantage of the algorithm in Prescott Adams and MacKay (2007) is that we can recursively compute the probability over the number of neighbours, $p(k_t)$, by keeping track of the joint probability of each k and the data: $p(k_{t-1}, x_0, ..., x_{t-1})$, as we observe a new datum, x_t alongside the predictive probability of x_t for a given number of neighbours, $\pi_t = p(x_t|k_{t-1}, \eta_{t-1})$.

$$p(k_t = k_{t-1} + 1, x_0, ..., x_t) = p(k_{t-1}, x_0, ..., x_{t-1})\, \pi_t\, p_\gamma \tag{4}$$

$$p(k_t = 0, x_0, ..., x_t) = \sum_{k_{t-1}} p(k_{t-1}, x_0, ..., x_{t-1})\, \pi_0\, (1 - p_\gamma) \tag{5}$$

Finally, we define the notation $\eta_\rho \nleftarrow x_t$ to indicate that we update the distribution parameters for η_ρ with the datum x_t using standard Bayesian updating rules[4] (see Fink (1997) for conjugate prior updating within the exponential family).

[1] We can use any valid distance metric to produce such ordered list.

[2] We note that the technique in Prescott Adams and MacKay (2007) does introduce a slight approximation error, evident mainly for short run lengths. Alas, computing the exact posterior would increase the complexity of the algorithm to $O(n^2)$.

[3] Covered in more detail in the implementation notes.

[4] As a simple example, let's assume we are updating the probability of a Bernoulli distribution, e.g. a coin toss, with a prior of $\alpha = 50$ for *heads* and $\beta = 50$ for *tails* (where $\eta = \{\alpha, \beta\}$ defined as a total of 100 pseudo-observations and a prior probability $p(H) = 0.5$). If we then observe a new datum $x = tails$, the $\eta \nleftarrow x$ operation will update the parameters to $\alpha' = 50$ and $\beta' = 51$, for a posterior predictive distribution of $p(H) = 0.49505$.

2.2.1 Implementation Notes

(a) The hazard function need not be constant; $p_\gamma = f()$. can, interestingly, depend on distance between points, or the current run-length (i.e. not memory-less), etc.;

Algorithm 1 Efficient bayesian k-NN algorithm

Initialize the data:
  $x_0, ..., x_{\tau-1} \leftarrow$ ordered data for target point τ
Initialize change-point variables:
  $p(k_0 = 0) \leftarrow 1.0$
  $\eta_0 \leftarrow \eta_{prior}$
for $t \leftarrow 0, \tau - 1$ **do**
  Observe next variable x_t
  for $i \leftarrow 0, t$ **do**
    Compute predictive probability:
      $\pi_i = p(x_t | k_{t-1} = i, \eta_i)$
    Compute growth probabilities:
      $p(k_t = k_{t-1} + 1, \vec{x}_{0:t}) = p(k_{t-1}, \vec{x}_{0:t-1})\, \pi_t\, p_\gamma$
  end for
  Compute change-point probability:
    $p(k_t = 0, \vec{x}_{0:t}) =$
      $\sum_{k_{t-1}} p(k_{t-1}, \vec{x}_{0:t-1})\, \pi_0\, (1 - p_\gamma)$
  Compute evidence:
    $p(\vec{x}_{0:t}) = \sum_{k_t} p(k_t, \vec{x}_{0:t})$
  for $i \leftarrow 0, t$ **do**
    Compute probability of k:
      $p(k_t = i | \vec{x}_{0:t}) = \frac{p(k_t = i, \vec{x}_{0:t})}{p(\vec{x}_{0:t})}$
    Update distributions:
      $\eta_i \twoheadleftarrow x_t$
  end for
end for
return $p(k_\tau | \vec{x}_{0:\tau-1}) \quad \forall k_\tau \in \{0, ..., \tau\}$

(b) Initializing the change-point probability: we do not have to set it to 1.0 before the first data-point. To speed up the analysis, we can start the algorithm m points away from our target point (where m can be set so that the prior probability of a change-point having occurred before m falls below a preset threshold, and $m \ll n$). In this case, and as an alternative, we can initialize the probability of a change-point before m with the prior distribution for p_γ.
(c) The Bayesian update defined as '$\twoheadleftarrow$' can generally be computed efficiently for distributions in the exponential family. Other, possibly more complex, distribution may require a quadrature or simulation approach.
(d) For large data-sets, we resort to applying a log transform the joint probabilities in order to maintain numerical stability.

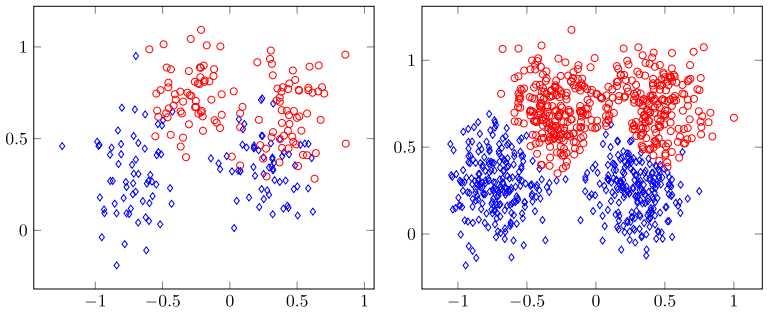

Fig. 3 Ripley's traning data (left) and test data (right)

3 Results

In order to critically appraise this approach, we benchmark our analysis to the ubiquitous Ripley data-set for classification (Fig. 3), and, as for a regression problem, to the Nuclear Power Plant output in Kaya et al. (2012). The comparison is made against results obtained using the *global* optimal number of neighbours (as a manual process). In other words, we compare this approach against the best choice of k when applied to all of the training data points. The key idea here is indeed that the optimal k varies depending on the specific target point within the same data-set.[5]

In addition to the a prediction based on various k values (weighted by their likelihood), we now have a measure of certainty regarding our prediction. In Fig. 4 we present the probability of classification computed using the training data.

The MCMC-based Bayesian analysis in Cucala et al. (2008) achieved a misclassification rate of 0.087, which is, not surprisingly, similar to our result. We also note the similarity between our Fig. 4 and the one presented in such study. Indeed, we are not proposing the idea of a Bayesian approach to estimating the number of neighbours, but an efficient method to do so. Using this algorithm, the time to compute the posterior distribution over k for a test point within the Ripley data-set averaged three milliseconds per test point (for a standard PC), compared to 50,000 paths used in the MCMC implementation in Cucala et al. (2008). Notably, our approach results in the *local* Bayesian analysis of k, i.e. specific to the data point being queried, as opposed to the global analysis for the MCMC approach.

Finally, on the right column in Table 1 we show how, by switching the prior from a Beta distribution in the classification problem with the Normal distribution (as we only assumed the mean to be unknown), we can obtain improved results for the Power Plant Output data. An interesting observation is that, if we plot the maximum posterior probability of the data w.r.t. the absolute error (in Fig. 5), we can observe the ultimate limitation of k-NN algorithms. Data points with a large absolute error have clearly a very small probability of occurring, despite the fact that we are displaying the *maximum* likelihood across all possible k's; in other words, these are true outliers with respect to the distance measure that we have chosen[6].

[5]Data and descriptions for both data-sets are available the at UCI Machine Learning Repository (http://www.ics.uci.edu)

[6]From an intuitive standpoint, we expect a plot of the maximum probability of the data across all values of k to present outliers as having very low probability; this would indicate that the point is indeed dissimilar to *any* grouping of neighbours for the chosen distance measure. In practice, a true outlier will have maximum probability for $k = 0$, i.e. when it belongs to the uninformed prior as its the distribution with the largest variance, hence large absolute errors in Fig. 5 converge onto a single line.

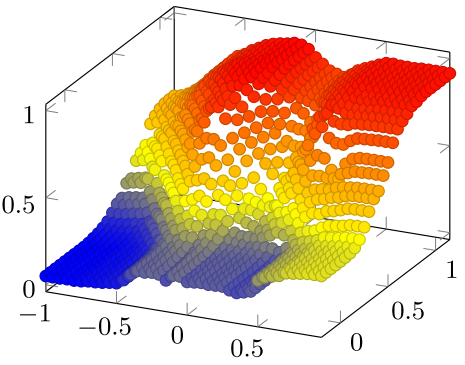

Fig. 4 Probability of classification using Ripley's traning data

Table 1 Comparison of the global optimal k neighbours versus the algorithm presented here: misclassification rate for Ripley data and average absolute error for Power Plant Output data

Algorithm	Ripley (Misclassification)	Power plant (Avg. Abs. Error)
k-NN (manual k search)	0.13	3.6
Bayesian k-NN	0.09	2.9

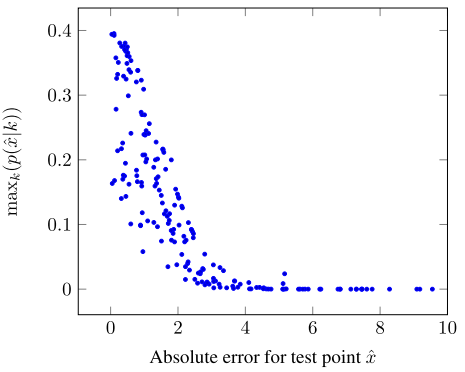

Fig. 5 Maximum probability for 200 samples of realized test data across all possible k neighbours (Power Plant Output data)

4 Conclusive Remarks

We have presented an efficient algorithm to compute a Bayesian analysis over the number of neighbours in k-NN algorithms, applicable to classification and regression, which does not rely on MCMC simulation. This yields both superior predictions and a full probabilistic view of k. Yet the biggest challenge for k-NN algorithms is likely to be within the choice of the distance measure and differentiated input scaling (as highlighted in Weinberger and Saul 2009). Certainly for multidimensional problems, the challenge lies in ordering the neighbours correctly with respect to their proximity to our target point, which in turn is driven by the coordinate transform we apply to compute the distance measure. An efficient Bayesian approach in understanding such scaling aspect may well be possible.

References

Cucala L, Marin J-M, Robert C, Titterington M (2008) A Bayesian reassessment of nearest-neighbour classification ArXiv e-prints

Fink D (1997) A compendium of conjugate priors

Ghosh AK (2006) On optimum choice of k in nearest neighbor classification. Comput Stat Data Anal 50(11):3113–3123

Green PJ (1995) Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika 82(4):711

Guo R, Chakraborty S (2010) Bayesian adaptive nearest neighbor. Stat Anal Data Min 3(2):92–105

Holmes CC, Adams NM (2002) A probabilistic nearest neighbour method for statistical pattern recognition. J Royal Stat Soc Ser B (Stat Methodol) 64(2):295–306

Ji WY, Friel N (2013) Efficient estimation of the number of neighbours in probabilistic K nearest neighbour classification. CoRR, arXiv:1305.1002

Kaya H, Tüfekci P, Gürgen FS (2012) Local and global learning methods for predicting power of a combined gas & steam turbine. In: International conference on emerging trends in computer and electronics engineering (ICETCEE 2012), Dubai

Manocha S, Girolami MA (2007) An empirical analysis of the probabilistic k-nearest neighbour classifier. Pattern Recogn Lett 28(13):1818–1824

Prescott Adams R, MacKay DJC (2007) Bayesian Online Changepoint Detection. ArXiv e-prints

Smith AFM (1975) A bayesian approach to inference about a change-point in a sequence of random variables. Biometrika 62(2):407–416

Stephens DA (1994) Bayesian retrospective multiple-changepoint identification. J Royal Stat Soc Ser C (Appl Stat) 43(1):159–178

Tomasev N, Radovanović M, Mladenić D, Ivanović M (2011) A probabilistic approach to nearest-neighbor classification: Naive hubness bayesian knn. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11. ACM, New York, pp 2173–2176

Weinberger KQ, Saul LK (2009) Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res 10:207–244