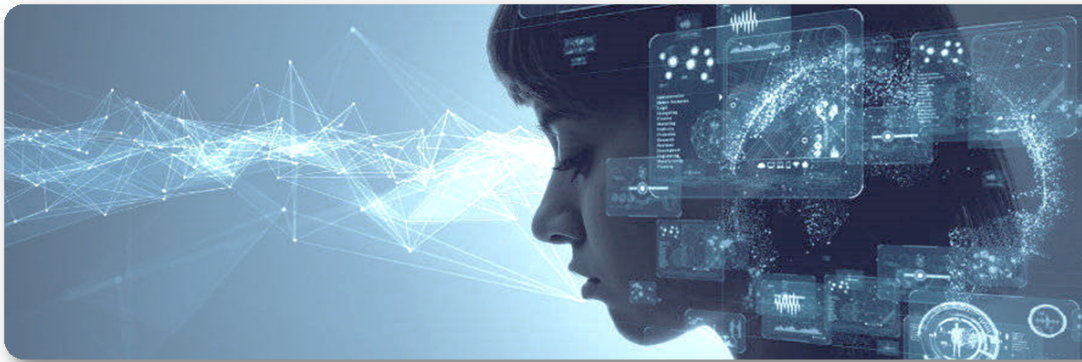

COMP 2211 Exploring Artificial Intelligence

K-Nearest Neighbors - Why $k = \sqrt{n}$?

Dr. Desmond Tsoi

Department of Computer Science & Engineering
The Hong Kong University of Science and Technology, Hong Kong SAR, China

Using the heuristic $k = \sqrt{n}$

- Adapting from Devroye, Gyorfi and Lugosi (1996), theoretical performance of the k-nearest neighbor classifier can be organized, albeit not exclusively, along the following lines:
  - k is determined to be a finite, fixed constant, whilst the sample size $n \to \infty$. The determination of k is a priori, i.e. selected in advance. That is, using prior knowledge, after exploratory data analysis, or using a heuristic.
  - $k \to \infty$ whilst $k/n \to 0$. Similar to above in a priori determination of k, but now k is not fixed relative to sample size n.
  - Data dependent methods for determining k e.g. using a training set, test set and selecting k to minimize the estimated classification error rate, or using cross-validation.

Using the heuristic $k = \sqrt{n}$

- The heuristic $k = \lfloor \sqrt{n} \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function, would fall under the second category. From the above book, the following is a quantitative, finite-sample probabilistic bound on the excess risk $L_n - L^*$, which in turn implies an asymptotic result:

Theorem 11.1 (Devroye and Gyorfi (1985), Zhao (1987))

Assume that each μ has a density. If $k \to \infty$ and $k/n \to 0$ then for every $\epsilon > 0$ there is an n_0 such that for $n > n_0$,

$$P(L_n - L^* > \epsilon) \leq 2e^{-n\epsilon^2/(72\gamma_d^2)},$$

where the γ_d is the minimal number of cones centered at the origin of angle $\pi/6$ that cover $\mathbb{R}$. (For the definition of a cone, see Chapter 5). Thus, the k-NN rule is strongly consistent.

Using the heuristic $k = \sqrt{n}$

- Supplying context on the terms not defined in the extract, $L_n = L_n(g_n) = P(g_n(X) \neq Y)$ is the risk of the k-nearest neighbor classifier $g_n(X)$, where g_n is estimated from a sample of size n.
- $L^* = L(g^x) = inf_{g \in \mathcal{G}} P(g(X) \neq Y)$, is the Bayes-optimal classification risk, or Bayes error rate, that is the risk of the Bayes classifier g^*. Glossing over the measure theoretic technicalities, that the measure μ has a density just means that X has a density.
- Parsing the theorem, the main condition requires that $k \to \infty$ as the sample size $n \to \infty$ in such a way that $k/n \to 0$. Your heuristic satisifies this condition because $k = \lfloor \sqrt{n} \rfloor \to \infty$ and $k/n = \lfloor \sqrt{n} \rfloor / n \approx (1/\sqrt{n}) \to 0$.
- Taking the theorem as an asymptotic result, the k-nearest neighbor classifier is strongly consistent, in the sense of

$$L_n \xrightarrow{\text{a.s.}} L^* \iff P\left(lim_{n \to \infty} L_n = L^*\right) = 1.$$

Using the heuristic $k = \sqrt{n}$

- That is, as you collect more observations $n \to \infty$, the classification error rate L_n of the k-nearest neighbor classifier g_n will converge almost surely to the minimal classification error rate you can possibly hope to achieve, L^*. And that this converges exponentially quickly.
- Furthermore, the result is non-asymptotic in that for finite n, it bounds the probability that L_n deviates from L^* by more than ϵ in terms of finite constants.
- **On the use of data-dependent methods for selecting k**
  The utility of the above theoretical result then is that it supplies insight on heuristics like the one you have outlined. Its limitations, like many results in statistical learning theory, is that the constant γ_d may be difficult to compute, or in the case that it is computable, renders the bound too loose to give any practical prescriptions.

Using the heuristic $k = \sqrt{n}$

- Echoing the sentiment expressed:

> k-nearest-neighbor density linkage is strongly set consistent for high-density (density-contour) clusters if k is chosen that $\frac{k}{n} \to 0$ and $\frac{k}{ln(n)} \to \infty$ as $n \to \infty$.

the authors advocate the use of data-dependent means of selecting k in practice:

> Consistency by itself may be obtained by choosing $k = \lfloor \sqrt{n} \rfloor$, but few "if any" users will want to blindly use such recipes. Instead, a healthy dose of feedback from the data is preferable.

- Similar consistency results for the use of a test set to select k based on minimising a holdout estimate of the classification error rate are supplied therein.
- There seems to be some consensus that the kind of result listed above is a continuation of a line of work in the spirit of Stone (1977). A more recent, specialised treatment is by Chaudhuri and Dasgupta (2014). Further details can be found in the references below.

References

- Devroye, L., Györfi, L., & Lugosi, G. (1996). A Probabilistic Theory of Pattern Recognition. Springer. https://doi.org/10.1007/978-1-4612-0711-5.Seechapters5,6,11,26.
- Stone, C. J. (1977). Consistent nonparametric regression. The Annals of Statistics, 5(4) 595 - 645. https://doi.org/10.1214/aos/1176343886
- Chaudhuri, K., & Dasgupta, S. (2014). Rates of convergence for nearest neighbour classification. Advances in Neural Information Processing Systems 27, NIPS 2014. https://papers.nips.cc/paper/2014/hash/db957c626a8cd7a27231adfbf51e20eb-Abstract.html

That's all!

Any questions?