Machine Learning (CE 40477)

Fall 2024

Ali Sharifi-Zarchi

CE Department
Sharif University of Technology

November 3, 2025

Parametric vs. Non-Parametric Methods

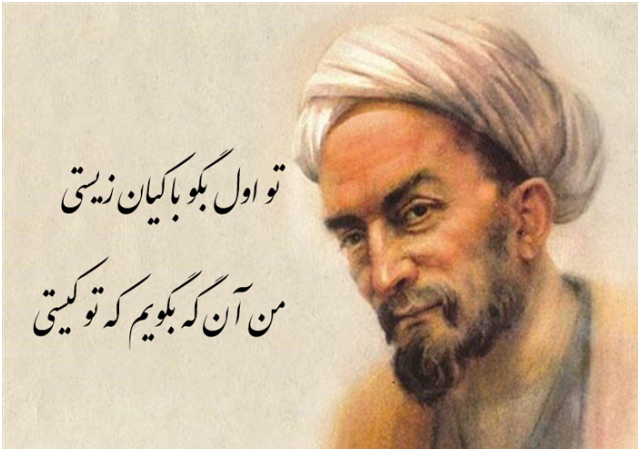
- **Parametric** methods **learn parameters** from data and then use these inferred parameters to make predictions on new data points
  - Learning: estimating parameters from data
  - e.g., **Linear Regression**, **Logistic Regression**
- **Non-parametric** methods
  - Training examples are used **directly**
  - A separate **training phase** is **not required**
  - e.g., **k-Nearest Neighbors (kNN)**
- Both supervised and unsupervised learning methods can be categorized as either parametric or non-parametric

Outline

Non-Parametric Approach

- **Unsupervised: Non-Parametric Density Estimation**
  - Parzen Windows
  - k-Nearest Neighbor Density Estimation
- **Supervised: Instance-Based Learners**
  - Classification
    - kNN Classification
    - Weighted (or Kernel) kNN
  - Regression
    - kNN Regression
    - Locally Linear Weighted Regression

kNN Concept



- First, tell me who you have lived with, and then I will tell you who you are

Figure adapted from https://setare.com

k-Nearest Neighbor (kNN)

- **kNN Classifier:** considers the $k \geq 1$ nearest neighbors of a query point
  - The label of x is predicted by **majority voting** among its k nearest neighbors
- Example: $k = 5$, $x = [x_1, x_2]$

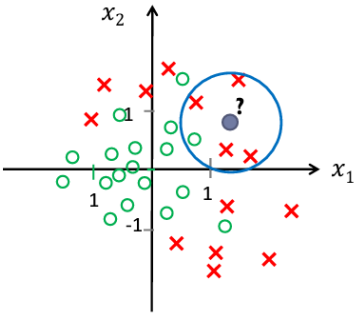


Figure adapted from M. Soleymani, ML slides, Sharif University of Technology

k-Nearest Neighbor (kNN) Classifier

- **Given:**
  - Training data $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$ are stored directly (no model parameters are learned)
- **To classify a new sample** x:
  - Find the k nearest training samples to x
  - Among these k samples, count how many k_j belong to each class C_j ($j = 1, \ldots, C$)
  - Assign x to the class C_{j^*}, where

$$j^* = \arg\max_{j=1,\ldots,C} k_j$$

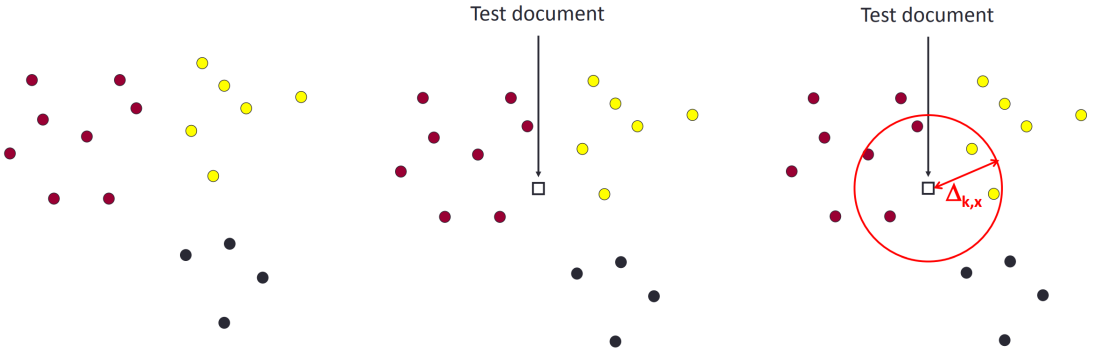
- The kNN algorithm is considered a **discriminative** method

k-Nearest Neighbor (kNN) Classifier

- The **kNN classifier** can produce **non-linear decision boundaries**, unlike previous methods such as linear and logistic regression
- However, this method can be quite sensitive to **outliers** or **noisy data**, particularly when:
  - The dataset is **small**
  - The data is **low-dimensional**
  - A **small value of** k is used — for instance, $k = 1$ depends only on the single nearest neighbor, which can be misleading in many test cases

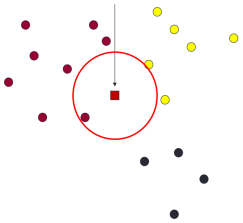
kNN Example

- We aim to classify a new document and assign it to one of three categories by examining its neighboring samples

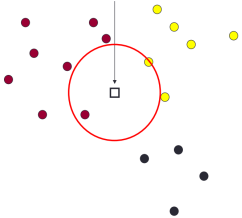


kNN Example

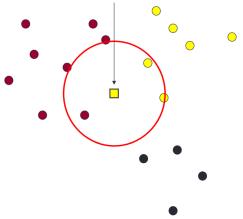
1-Nearest Neighbor



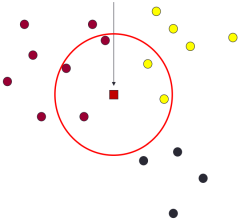
2-Nearest Neighbors



3-Nearest Neighbors

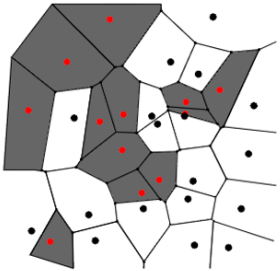


5-Nearest Neighbors



Voronoi Tessellation

- **Voronoi Tessellation:**
  - Each cell contains all points that are closer to a given training sample than to any other sample
  - All points within a cell share the label of the corresponding training sample



Voronoi Tessellation

- The 1-NN decision regions form a Voronoi tessellation

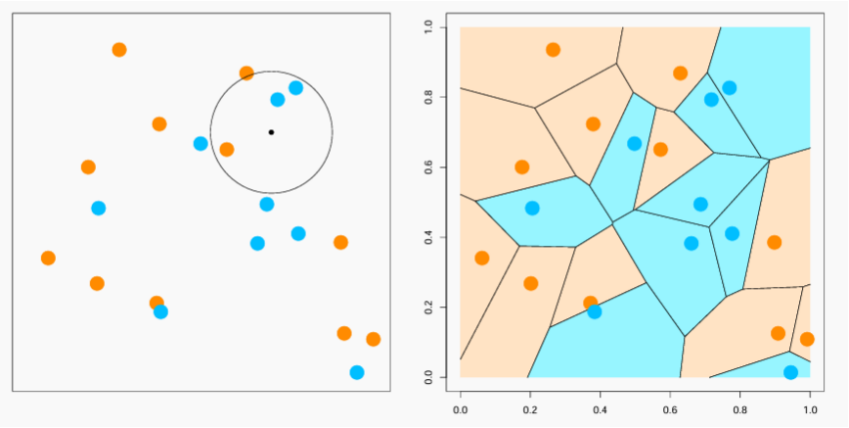
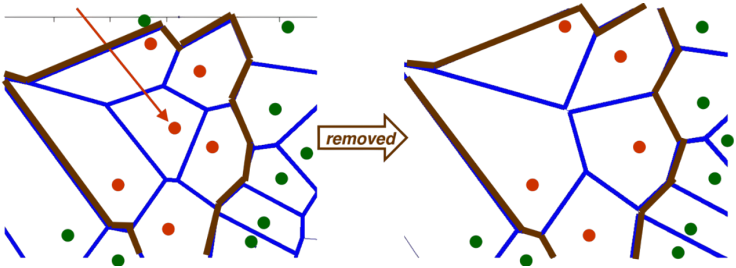


Figure adapted from R. Zhu, Stat 542: Statistical Learning k-Nearest Neighbor and the BiasVariance Trade-Off, Lecture Notes

Reducing Complexity: Editing 1-NN

- If all Voronoi neighbors of a sample belong to the same class, that sample is redundant and can be safely removed



- The number of stored samples decreases
- The decision boundaries are guaranteed to remain unchanged

Nearest Neighbor Classifier: Error Bound

- **Nearest Neighbor:** k-NN with $k = 1$
  - Decision rule:

$$\hat{y} = y^{NN(\mathbf{x})} \quad \text{where} \quad NN(\mathbf{x}) = \arg\min_{i=1,\dots,N} \left\| \mathbf{x} - \mathbf{x}^{(i)} \right\|$$

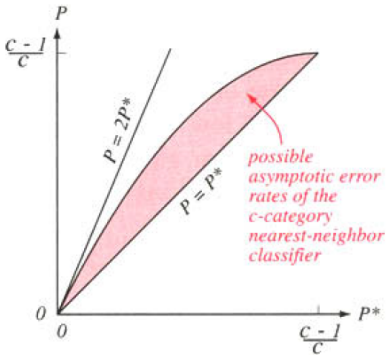
- **Cover & Hart (1967):** The asymptotic risk of the NN classifier satisfies

$$R^* \leq R_\infty^{NN} \leq 2R^*(1 - R^*) \leq 2R^*$$

R_n: expected risk of the NN classifier with n training examples drawn from $p(\mathbf{x}, y)$

$R_\infty^{NN} = \lim_{n\to\infty} R_n^{NN}$

R^*: the optimal Bayes risk



Cover & Hart (1967)

Nearest Neighbor Pattern Classification

T. M. COVER, MEMBER, IEEE, AND P. E. HART, MEMBER, IEEE

***Abstract*—The nearest neighbor decision rule assigns to an unclassified sample point the classification of the nearest of a set of previously classified points. This rule is independent of the underlying joint distribution on the sample points and their classifications, and hence the probability of error R of such a rule must be at least as great as the Bayes probability of error R^*—the minimum probability of error over all decision rules taking underlying probability structure into account. However, in a large sample analysis, we will show in the M-category case that $R^* \leq R \leq R^*(2 - MR^*/(M-1))$, where these bounds are the tightest possible, for all suitably smooth underlying distributions. Thus for any number of categories, the probability of error of the nearest neighbor rule is bounded above by twice the Bayes probability of error. In this sense, it may be said that half the classification information in an infinite sample set is contained in the nearest neighbor.**

I. Introduction

IN THE CLASSIFICATION problem there are two extremes of knowledge which the statistician may possess. Either he may have complete statistical knowledge of the underlying joint distribution of the observation x and the true category θ, or he may have no knowledge of the underlying distribution except that which can be inferred from samples. In the first extreme, a standard Bayes analysis will yield an optimal decision procedure and the corresponding minimum (Bayes) probability of error of classification R^*. In the other extreme, a decision to classify x into category θ is allowed to depend only on a collection of n correctly classified samples $(x_1, \theta_1), (x_2, \theta_2), \cdots, (x_n, \theta_n)$, and the decision procedure is by no means clear. This problem is in the domain of nonparametric statistics and no optimal classification procedure exists with respect to all underlying statistics.

If it is assumed that the classified samples (x_i, θ_i) are *independently identically distributed* according to the distribution of (x, θ), certain heuristic arguments may be made about good decision procedures. For example, it is reasonable to assume that observations which are close together (in some appropriate metric) will have the same classification, or at least will have almost the same posterior probability distributions on their respective classifications. Thus to classify the unknown sample x we may wish to weight the evidence of the nearby x_i's most heavily. Perhaps the simplest nonparametric decision procedure of this form is the *nearest neighbor* (NN) rule, which classifies x in the category of its nearest neighbor. Surprisingly, it will be shown that, in the large sample case, this simple rule has a probability of error which

Manuscript received February 23, 1966; revised April 29, 1966. This work has been supported at Stanford University by U. S. Army Electronics Command under Contract DA28-043-AMC-01764(E) and by USAF under Contract AF49(638)1517; and at the Stanford Research Institute, Menlo Park, Calif., by RADC under Contract AF30(602)-3945.

T. M. Cover is with the Department of Electrical Engineering, Stanford University, Stanford, Calif.

P. E. Hart is with the Stanford Research Institute, Menlo Park, Calif.

isl.stanford.edu/ cover/papers/transIT/0021cove.pdf

Bayes Optimal Classifier

Assume (although this is almost never the case) that we know $P(y|\mathbf{x})$ Then, we can simply predict the most likely label

$$\text{The Bayes optimal classifier predicts:} \quad y^* = \arg\max_y P(y|\mathbf{x})$$

Although the Bayes optimal classifier performs as well as possible, it can still make mistakes It is wrong whenever a sample does not belong to the most likely class We can compute the probability of that happening exactly, which gives the error rate

$$\epsilon_{\text{BayesOpt}} = 1 - \mathbb{P}(y^*(\mathbf{x})|\mathbf{x}) = 1 - \mathbb{P}(y^*|\mathbf{x})$$

Bayes Optimal Classifier: Example

Assume, for example, that an email $\mathbf{x}$ can be classified as either spam $(+1)$ or ham (-1)
For the same email $\mathbf{x}$, the conditional class probabilities are

$$P(+1|\mathbf{x}) = 0.8 \qquad P(-1|\mathbf{x}) = 0.2$$

In this case, the Bayes optimal classifier would predict the label

$$y^* = +1$$

as it is the most likely one, and its error rate would be

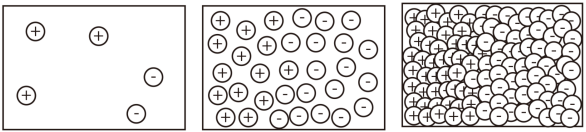
$$\epsilon_{\text{BayesOpt}} = 0.2$$

Why is the Bayes optimal classifier interesting if it cannot be used in practice The reason is that it provides a highly informative lower bound on the error rate Given the same feature representation, no classifier can achieve a lower error We use this fact to analyze the error rate of the k-NN classifier

1-NN Convergence Proof

Cover and Hart (1967): As $n \to \infty$, the 1-NN error is no more than twice the error of the Bayes optimal classifier (Similar guarantees also hold for $k > 1$)

n small $\quad$ n large $\quad$ $n \to \infty$



- As $n \to \infty$, each test point x_t becomes almost identical to its nearest neighbor x_{NN}, so

$$P(y^*|x_t) \approx P(y^*|x_{NN})$$

- An error occurs only when their labels differ, with probability

$$2P(y^*|x_t)(1 - P(y^*|x_t)) = 2\epsilon_{\text{BayesOpt}}$$

- Thus, the 1-NN error is **at most twice** the Bayes optimal error

1-NN Convergence Proof

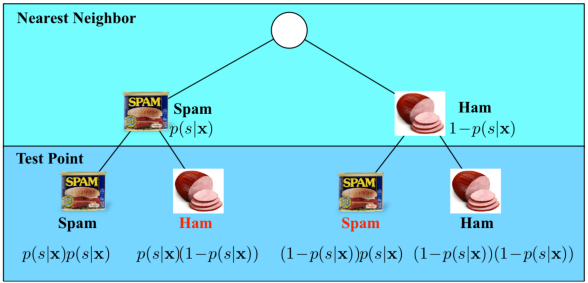
Let $\mathbf{x}_{NN}$ be the nearest neighbor of the test point $\mathbf{x}_t$ As $n \to \infty$, the distance $\text{dist}(\mathbf{x}_{NN}, \mathbf{x}_t)$ approaches zero

$$\begin{aligned}\epsilon_{NN} &= \mathbb{P}(y^*|\mathbf{x}_t)(1-\mathbb{P}(y^*|\mathbf{x}_{NN})) + \mathbb{P}(y^*|\mathbf{x}_{NN})(1-\mathbb{P}(y^*|\mathbf{x}_t)) \\ &\leq (1-\mathbb{P}(y^*|\mathbf{x}_{NN})) + (1-\mathbb{P}(y^*|\mathbf{x}_t)) \\ &= 2(1-\mathbb{P}(y^*|\mathbf{x}_t)) = 2\epsilon_{\text{BayesOpt}}\end{aligned}$$

where the inequality follows from $\mathbb{P}(y^*|\mathbf{x}_t) \leq 1$, and we assume that $\mathbb{P}(y^*|\mathbf{x}_t) = \mathbb{P}(y^*|\mathbf{x}_{NN})$

1-NN Convergence Proof

Possible labels: Spam, Ham (non-spam email)



In the limiting case, the test point and its nearest neighbor are identical There are exactly two cases in which a misclassification can occur — when the test point and its nearest neighbor have different labels The probability of this event is given by the sum of the two red regions:

$$(1-p(s|\mathbf{x}))p(s|\mathbf{x}) + p(s|\mathbf{x})(1-p(s|\mathbf{x})) = 2p(s|\mathbf{x})(1-p(s|\mathbf{x}))$$

Curse of Dimensionality

Good News: As $n \to \infty$, the 1-NN classifier performs at most twice as poorly as the best possible classifier

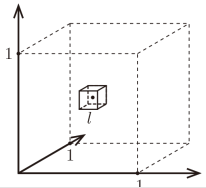
Bad News: We are cursed

Idea: The kNN classifier assumes that similar points have similar labels However, in high-dimensional spaces, randomly drawn points are rarely close to each other

Pitfalls: The Curse of Dimensionality

- Low-dimensional visualizations can be misleading. In high-dimensional spaces, most points are far apart
- If we want the nearest neighbor to be closer than ϵ, how many points are required to guarantee this
- The volume of a single ball with radius ϵ is $\mathcal{O}(\epsilon^d)$
- The total volume of $[0,1]^d$ is 1
- Therefore, $\mathcal{O}\left(\left(\frac{1}{\epsilon}\right)^d\right)$ balls are needed to cover the entire volume

Example: Points are uniformly sampled within the unit cube $[0,1]^d$ Let ℓ be the edge length of the smallest hypercube that contains all k nearest neighbors of a test point



Curse of Dimensionality

$$\ell^d \approx \frac{k}{n} \quad \Rightarrow \quad \ell \approx \left(\frac{k}{n}\right)^{1/d}$$

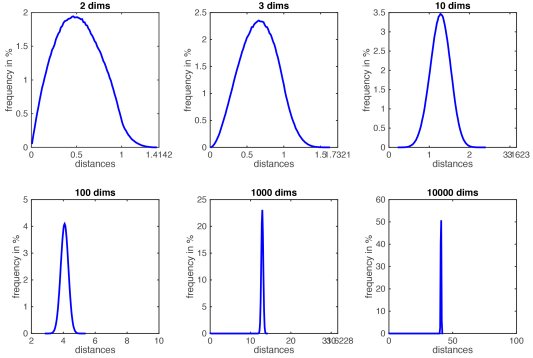
For $n = 1000$ and $k = 10$, we have:

d	ℓ
2	0.10
10	0.63
100	0.955
1000	0.9954

Observation: As d increases, almost the entire space is needed to find the k nearest neighbors, breaking the fundamental assumption of kNN

Curse of Dimensionality

Distance Distributions: The histograms below show the pairwise distances between random points in d-dimensional unit cubes As d increases, these distances concentrate within a very narrow range, meaning that almost all points are equally far apart



Curse of Dimensionality

Consequence: In high-dimensional spaces, the nearest neighbors are not truly close or similar This breaks the fundamental assumption underlying kNN

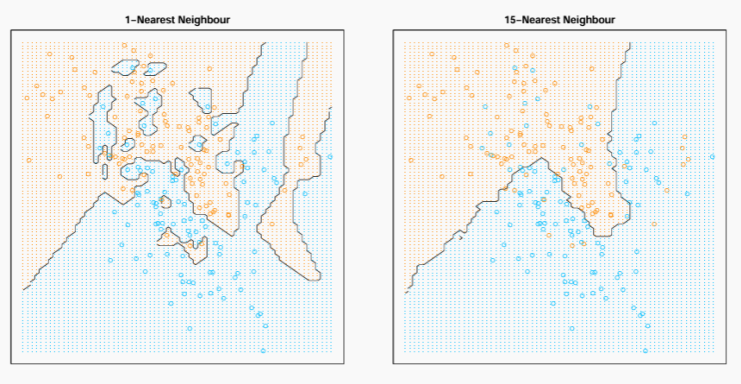
Data Growth Requirement: To make ℓ small, we must increase n:

$$\ell = \frac{1}{10} \Rightarrow n = \frac{k}{\ell^d} = k \times 10^d$$

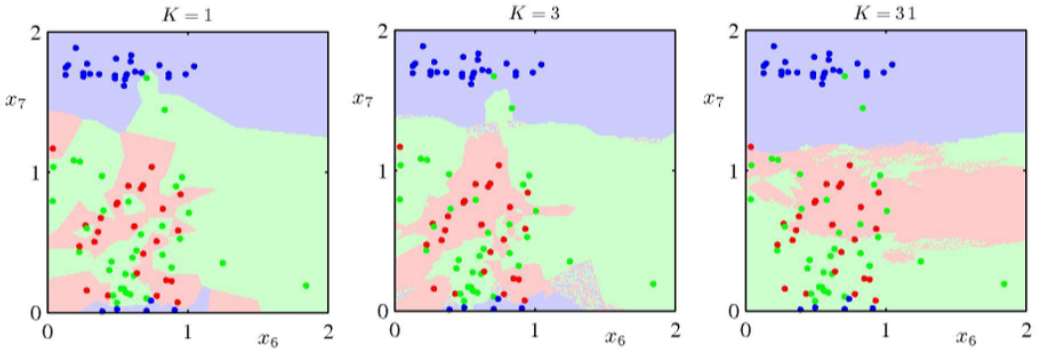
This quantity grows exponentially with d For $d > 100$, we would need more data points than there are electrons in the universe

Effect of k

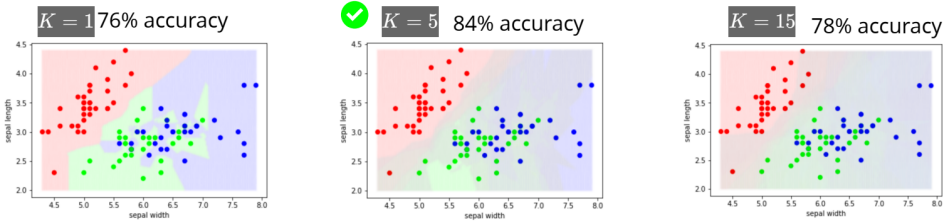
- Compare the results for $k = 1$ and $k = 15$
- As we increase k, the model becomes smoother and less complex



Effect of k

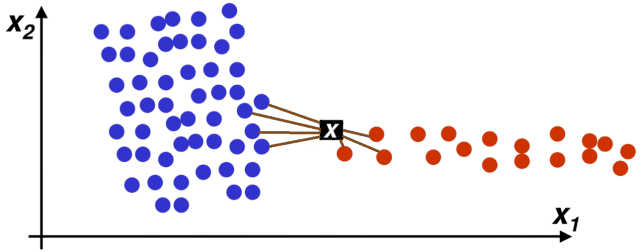


Choice of k



- In practice:
  - k should be large enough to minimize the overall error rate
    - A very small k produces noisy and unstable decision boundaries
  - k should also be small enough to ensure that only nearby samples are taken into account
    - A very large k produces overly smooth boundaries that may ignore local patterns
- Balancing these two requirements is not trivial
  - This is a common challenge in machine learning — we need to smooth data, but not too much

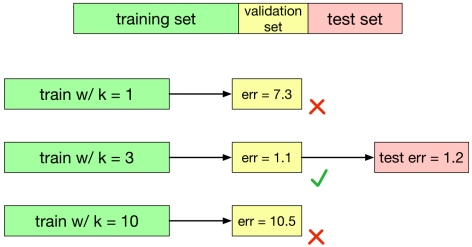
Choice of k: Example



- For $k = 1, \ldots, 7$, the point $\mathbf{x}$ is correctly classified as belonging to the red class
- For larger values of k, the classification of $\mathbf{x}$ becomes incorrect and it is assigned to the blue class
- BiasVariance Tradeoff in kNN:
  - Larger $k \Rightarrow$ Underfitting (high bias, low variance)
  - Smaller $k \Rightarrow$ Overfitting (low bias, high variance)

Validation and Test Sets

- k is an example of a hyperparameter — a parameter that cannot be directly learned during the training process
- We can tune hyperparameters by evaluating model performance on a separate validation set



- The test set is used only once, at the very end, to evaluate the generalization performance of the final trained model

Computational Complexity

What is the computational complexity of the kNN algorithm

- The basic kNN algorithm stores all training examples. Suppose we have n examples, each with d features
  - $O(d)$ to compute the distance to a single example
  - $O(nd)$ to find the nearest neighbor among all samples
  - $O(knd)$ to find the k closest examples
  - Thus, the total computational complexity is $O(knd)$
- This becomes prohibitively expensive when the number of samples is large
- However, kNN requires a large dataset to perform well

How can we make this process more efficient

Reducing Complexity: kNN Prototypes

Instead of storing every individual data point, we can group similar samples together and represent each group by a prototype These prototypes are organized into search trees, where:

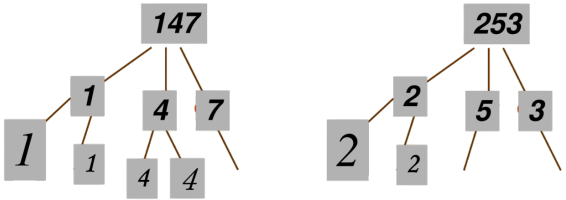
- Each node represents a cluster of similar data points
- The root node summarizes several prototypes or subgroups

During classification, we only search within the most relevant branches, not the entire dataset

- **Advantages:** The computational complexity decreases
- **Disadvantages:**
  - Finding a good search tree is not trivial
  - It may not always find the true nearest neighbor, and therefore it is **not** guaranteed that the decision boundaries remain unchanged

kNN Prototypes: Example

- The tree on the left represents class 1, and the tree on the right represents class 2
- Each gray box (e.g., 147, 253) denotes a prototype node that summarizes several samples
- The child nodes can be either:
  - Other prototypes representing subgroups, or
  - Individual samples
- In this way, the data are organized hierarchically to enable efficient search



Advantages and Disadvantages of kNN

- **Advantages**
  - Can be applied to data from any distribution
  - Very simple and intuitive to implement
  - Provides good classification performance if the number of samples is sufficiently large
- **Disadvantages**
  - Selecting the best value of k can be challenging
  - Computationally expensive, although certain optimizations are possible
  - Requires a large number of samples to achieve high accuracy
    - This limitation cannot be overcome without assuming a parametric distribution

Instance-Based Learner

- Main components required to construct an instance-based learner:
  - A distance metric to measure similarity between data points
  - The number of nearest neighbors of the test instance to consider
  - A weighting function (optional)
  - A method to determine the output based on the neighbors

Distance Measures

- **Minkowski Distance**

$$d(x, x') = \left(\sum_{i=1}^{n} |x_i - x'_i|^p \right)^{\frac{1}{p}}$$

  - For $p \geq 1$, this is a valid distance metric
  - When $p = 2$, the Minkowski distance is equivalent to the Euclidean distance
  - The Minkowski distance is also equal to the L^p norm of $(x - x')$

- Recall the definition of the L^p norm from linear algebra:

$$\|x\|_p = \sqrt[p]{|x_1|^p + \cdots + |x_n|^p}$$

$$\text{Some common } L^p \text{ norms: } \begin{cases} \|x\|_1 & = \sum_{i=1}^{n} |x_i| \\ \|x\|_2 & = \sqrt{x_1^2 + \cdots + x_n^2} \\ \|x\|_\infty & = \max\{|x_1|, |x_2|, \ldots, |x_n|\} \end{cases}$$

Distance Measures

- **Euclidean Distance**

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + \cdots + (x_d - x'_d)^2}$$

- **Distance Learning Methods:**
  - **Weighted Euclidean Distance**

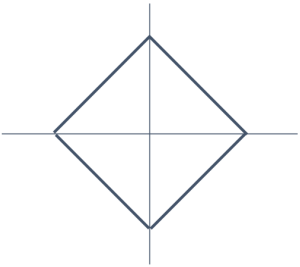
$$d_w(x, x') = \sqrt{w_1(x_1 - x'_1)^2 + \cdots + w_d(x_d - x'_d)^2}$$

  - Large weight $w_i \Rightarrow$ Attribute i is more important
  - Small weight $w_i \Rightarrow$ Attribute i is less important
  - Zero weight $w_i \Rightarrow$ Attribute i does not contribute to the distance

Distance Measures

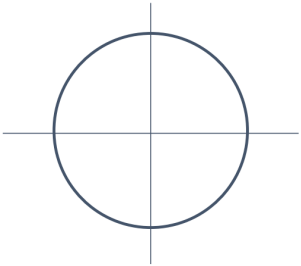
L_1 (Manhattan) Distance

$$d_1(I_1, I_2) = \sum_p \left| I_1^p - I_2^p \right|$$



L_2 (Euclidean) Distance

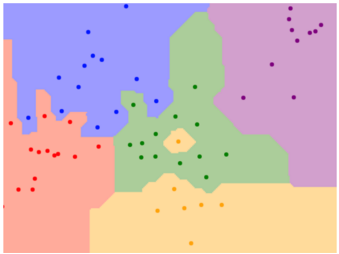
$$d_2(I_1, I_2) = \left(\sum_p (I_1^p - I_2^p)^2 \right)^{\frac{1}{2}}$$



Distance Measures ($k = 1$)

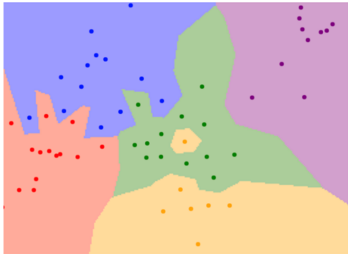
L_1 (Manhattan) Distance

$$d_1(I_1, I_2) = \sum_p \left| I_1^p - I_2^p \right|$$



L_2 (Euclidean) Distance

$$d_2(I_1, I_2) = \left(\sum_p (I_1^p - I_2^p)^2 \right)^{\frac{1}{2}}$$

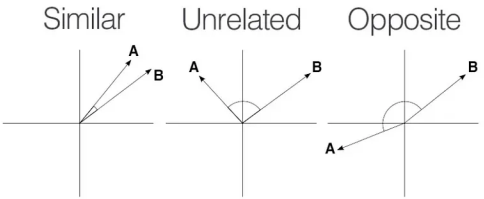


Distance Measures

- **Cosine Distance (Angle-Based Similarity)**

$$d(x, x') = 1 - \text{cosine similarity}(x, x')$$

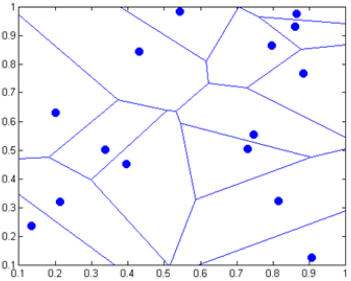
$$\text{where} \quad \text{cosine similarity}(x, x') = \frac{x \cdot x'}{\|x\|_2 \, \|x'\|_2} = \frac{\sum_{i=1}^{d} x_i x'_i}{\sqrt{\sum_{i=1}^{d} x_i^2} \sqrt{\sum_{i=1}^{d} (x'_i)^2}}$$



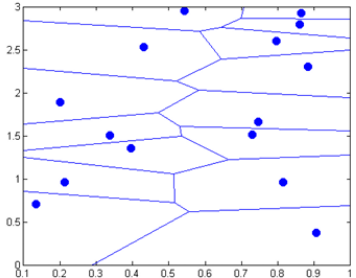
Example of angular difference in cosine similarity

Figure adapted from https://medium.com/@milana.shxanukova15/cosine-distance-and-cosine-similarity-a5da0e4d9ded

Effect of Distance Measure



$$d(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2}$$



$$d(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{(x_1 - x_1')^2 + 3(x_2 - x_2')^2}$$

Effect of different distance measures on classification boundaries

Figures adapted from M. Soleymani, ML slides, Sharif University of Technology

Effect of Distance Measure

You can get quite creative with the choice of the distance function For example, using *bending energy*, which measures how much transformation is required to make one example resemble another

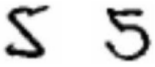


Fig. 1. Examples of two handwritten digits. In terms of pixel-to-pixel comparisons, these two images are quite different, but to the human observer, the shapes appear to be similar.

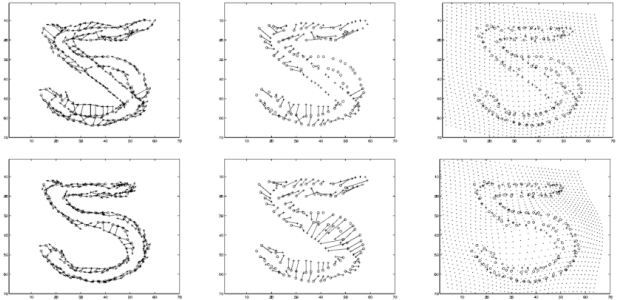


Fig. 4. Illustration of the matching process applied to the example of Fig. 1. Top row: 1st iteration. Bottom row: 5th iteration. Left column: estimated correspondences shown relative to the transformed model, with tangent vectors shown. Middle column: estimated correspondences shown relative to the untransformed model. Right column: result of transforming the model based on the current correspondences; this is the input to the next iteration. The grid points illustrate the interpolated transformation over $\mathbf{R}^2$. Here, we have used a regularized TPS model with $\lambda_0 = 1$.

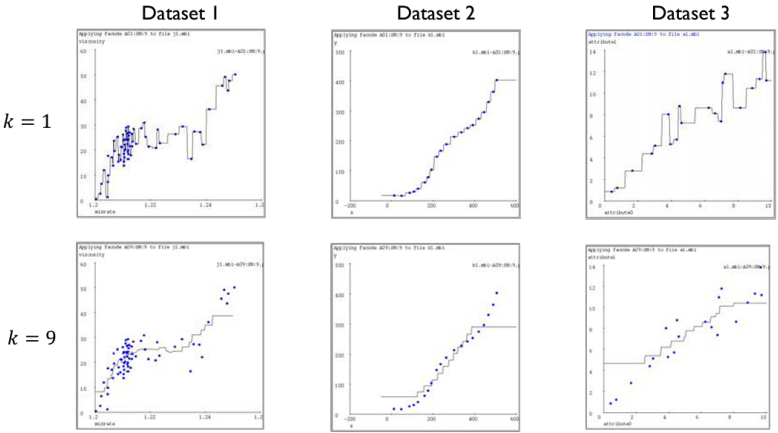
kNN Regression

- Let $x'^{(1)}, \ldots, x'^{(k)}$ be the k nearest neighbors of x, and let $y'^{(1)}, \ldots, y'^{(k)}$ be their corresponding labels

$$\hat{y} = \frac{1}{k} \sum_{j=1}^{k} y'^{(j)}$$

- Some challenges of kNN regression when fitting functions:
  - Discontinuities in the estimated function
  - For 1-NN: prone to noise fitting
  - For $k > 1$: averaging smooths out noise but may introduce other issues (e.g., flattening near the boundaries)

kNN Regression



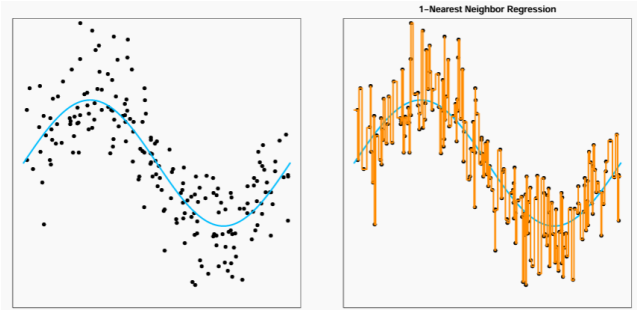
Figures adapted from Andrew Moores tutorial on Instance-Based Learning

kNN Regression: Example

- Suppose we have a dataset with a single feature uniformly distributed over $[0, 2\pi]$
- The true model is defined as:

$$Y = 2\sin(X) + \epsilon$$

- Where ϵ is a standard normal error term
- We simulate 200 observations and visualize the model for $k = 1$



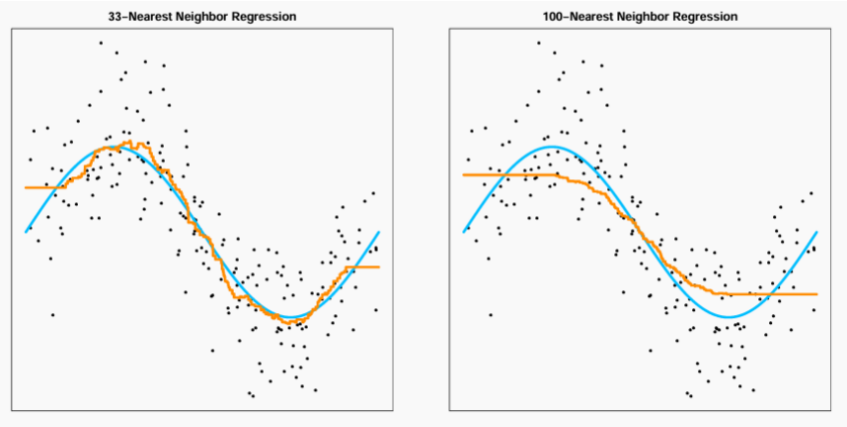
kNN Regression: Example

- Results for $k = 2$ and $k = 10$

Figures adapted from Andrew Moores tutorial on Instance-Based Learning

kNN Regression: Example

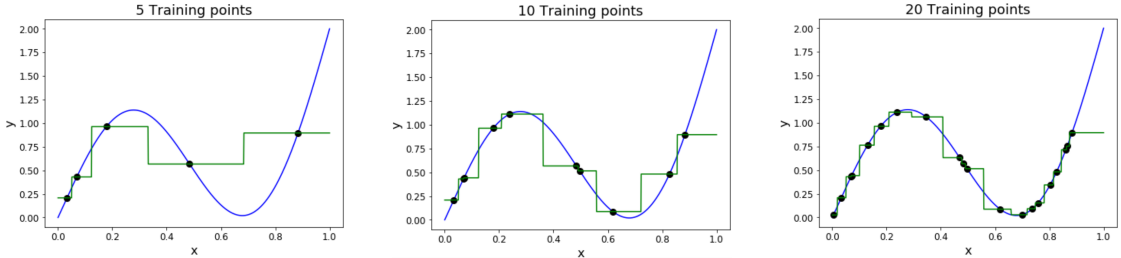
- As k increases, the model becomes smoother. However, if k is too large, it starts to deviate from the true underlying function



Figures adapted from Andrew Moores tutorial on Instance-Based Learning

kNN: Universal Approximation

As the number of training samples approaches infinity, the nearest neighbor method can approximate any function*!



The kNN regression fits a piecewise constant function (green), which approximates the true curve (blue) by averaging the nearby training samples (black)

* Subject to several technical conditions: only continuous functions on a compact domain, with assumptions on the spacing of training points, etc

Locally Weighted Linear Regression

- For each test sample, it produces a linear approximation of the target function in a local region
- Instead of computing the output by weighted averaging (as in kernel regression), we fit a parametric function locally

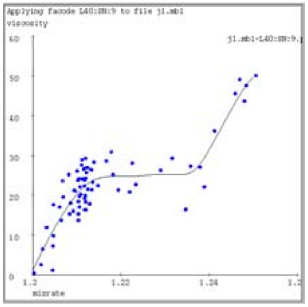
$$\hat{y} = f\left(x, x^{(1)}, y^{(1)}, \ldots, x^{(n)}, y^{(n)}\right)$$

$$\hat{y} = f(x; \mathbf{w}) = w_0 + w_1 x_1 + \cdots + w_d x_d$$

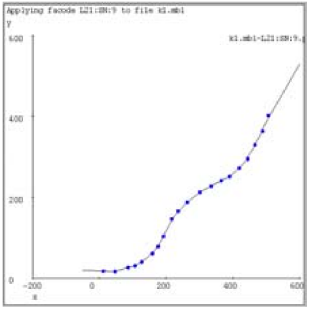
$$J(\mathbf{w}) = \sum_{i \in \mathcal{N}_k(x)} \left(y^{(i)} - \mathbf{w}^T x^{(i)}\right)^2$$

The weight vector $\mathbf{w}$ is computed separately for each test sample

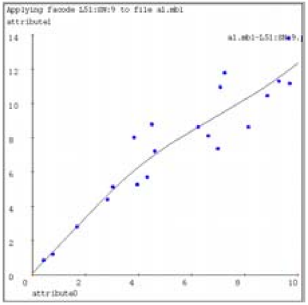
Locally Weighted Linear Regression: Example



$\sigma = \frac{1}{32}$ of x-axis width



$\sigma = \frac{1}{16}$ of x-axis width



$\sigma = \frac{1}{8}$ of x-axis width

- Produces smoother and more accurate results than weighted kNN regression

Vizier

Vizier: Old but fast locally weighted regression for Windows

Authors: Artur Dubrawski, Mike Duggan, Mary Soon Lee, Andrew Moore, Jeff Schneider, David Stott

Vizier was written in the mid-90's and was a combination of fast algorithms for the following operations, with a Graphical User Interface driver for Windows:

- K Nearest neighbor classification
- K Nearest neighbor regression
- Kernel regression (aka Shepherd's Interpolation)
- Kernel classification
- Locally weighted linear regression
- Locally weighted polynomial regression
- Locally weighted Bayesian polynomial regression
- Automatic determination of best model and distance metric

These days we have faster algorithms, and they run under Windows and Unix, but the functionality is not quite the same as the wild world of Vizier, so we are puttin might find it interesting or useful. Useful Links:

- Download Vizier
- A tutorial on using Vizier.
- A paper on some of the fast algorithms in Vizier
- Tutorial slides on locally weighted learning algorithms
- More recent, faster, Auton Lab local regression code: Bug Andrew About This! It's ready for applicification but just needs a final piece of work by Andrew!
- If you have questions or comments, please contact Andrew Moore

www.cs.cmu.edu/~awm/vizier/

Contributions

- **These slides are authored by:**
  - Danial Gharib
  - Mahan Bayhaghi
  - Erfan Jafari
  - Mahdi Aghaei

[1] C. M., *Pattern Recognition and Machine Learning*.
Information Science and Statistics, New York, NY: Springer, 1 ed., Aug. 2006.

[2] M. Soleymani Baghshah, "Machine learning." Lecture slides.

[3] M. Soleymani Baghshah, "Modern information retrieval." Lecture slides.

[4] T. Mitchell, *Machine Learning*.
McGraw-Hill series in computer science, New York, NY: McGraw-Hill Professional, Mar. 1997.

[5] R. Zhu, "Stat 542: Statistical learning - k-nearest neighbor and the bias-variance trade-off." Lecture notes.

[6] E. Xing, "Theory of classification and nonparametric classifier." Lecture notes.