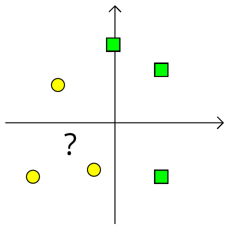
k-Nearest Neighbors and the Curse of Dimensionality

Cornell CS 3/5780 · Spring 2026

0. No Free Lunch

- Every ML algorithm must make assumptions!
- Choice of algorithm encodes assumptions about data set/distribution
- There is no one perfect approach for all problems!



- Common assumption: the relationship between $\mathbf{x}$ and y is locally smooth

1. The k-NN Algorithm

- **Core Assumption**: Similar inputs have similar outputs.
- **Classification Rule**: For a test input $\mathbf{x}$, assign the most common label amongst its k most similar training inputs.
- **Formally**: Let $S_{\mathbf{x}} \subseteq D$ be the set of k neighbors such that

$$\text{for all } (\mathbf{x}', y') \in D \setminus S_{\mathbf{x}}, \qquad \text{dist}(\mathbf{x}, \mathbf{x}') \geq \max_{(\mathbf{x}'', y'') \in S_{\mathbf{x}}} \text{dist}(\mathbf{x}, \mathbf{x}'')$$

  then the prediction is given by

$$h(\mathbf{x}) = \text{mode}(y'' : (\mathbf{x}'', y'') \in S_{\mathbf{x}})$$

- **Tie-Breaking Tip**: In case of a draw, return the result of k-NN with a smaller k.
- **Question:** What happens when $k = 1$? $k = |D| = n$?

2. Distance Metrics

The classifier fundamentally relies on a distance metric; the better it reflects label similarity, the better the classifier.

Minkowski Distance

$$\text{dist}(\mathbf{x}, \mathbf{x}') = \left(\sum_{r=1}^{d} |x_r - x'_r|^p \right)^{1/p}$$

- **Question:** what is the Minkowski Distance for:
  - $p = 1$
  - $p = 2$
  - $p \to \infty$

3. Constant Classifier

- **Concept**: Predicting the same label independent of the features.
- **Question**: What is the best constant classifier?
- **Significance**: Provides a baseline for debugging. Your classifier should perform much better!

4. Bayes Optimal Classifier

- **Concept**: Predicting the most likely label if you knew the conditional distribution $P(y|\mathbf{x})$.
- **Prediction**: $y^* = h_{\text{opt}}(\mathbf{x}) = \text{argmax}_y\, P(y|\mathbf{x})$.
- **Error Rate**: $\epsilon_{\text{BayesOpt}} = 1 - P(y^*|\mathbf{x})$.
- **Significance**: Provides a theoretical lower bound on the achievable error rate.

5. 1-NN Convergence Proof

- **Theorem (Cover and Hart, 1967)**: As $n \to \infty$, the 1-NN error for binary classification is no more than twice the Bayes error.
- **Key Mechanism**: As $n \to \infty$, the distance to the nearest neighbor $\text{dist}(\mathbf{x}_{NN}, \mathbf{x}_t) \to 0$, making $\mathbf{x}_{NN}$ identical to $\mathbf{x}_t$.
- **Proof Idea**: What is the probability that the label of $\mathbf{x}_{\text{NN}}$ is not the label of $\mathbf{x}_{\text{t}}$?
- **Question**: Explain each of the following steps
  - $\epsilon_{NN} = \text{P}(y^*|\mathbf{x}_t)(1 - \text{P}(y^*|\mathbf{x}_{NN})) + \text{P}(y^*|\mathbf{x}_{NN})(1 - \text{P}(y^*|\mathbf{x}_t))$
  - $\leq (1 - \text{P}(y^*|\mathbf{x}_{NN})) + (1 - \text{P}(y^*|\mathbf{x}_t))$
  - $= 2(1 - \text{P}(y^*|\mathbf{x}_t))$
  - $= 2\epsilon_{\text{BayesOpt}}$

6. The Curse of Dimensionality

- Points drawn from a probability distribution tend to never be close together in high dimensions.
- **Volume Analysis**: For uniform distribution on features, to capture k neighbors in a unit cube $[0,1]^d$, the required edge length $\ell^d \approx k/n$
- **Question**:
  - What happens to ℓ for k/n fixed and d getting big?
  - How big does n need to get to keep ℓ constant?

7. Mitigating The Curse

- **Linear Separation**: Pairwise distances between points grow with dimensionality, but distances to hyperplanes do not.
- **Low Dimensional Structure**: Data often lies on low-dimensional manifolds despite a high-dimensional d.
- **Question**: Images of faces have low dimensional structure. Why?

8. Summary of kNN

- Simple and effective classifier if distances reliably correspond to meaningful notion of dissimilarity.
- Provably accurate as $n \to \infty$, but also becomes slow.
- For large d, "neighbors" may no longer be similar to each other, so the key assumption breaks down