ONLINE MASTERS IN **DATA SCIENCE**

DSC 255 - MACHINE LEARNING FUNDAMENTALS

CROSS-VALIDATION

SANJOY DASGUPTA, PROFESSOR

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

K-nearest neighbor classification

Classify a point using the labels of its k-nearest neighbors among the training points.

MNIST:

K	1	3	5	7	9	11
Test error (%)	3.09	2.94	3.13	3.10	3.43	3.34

In real life, there’s no test set. How to decide which k is best?

K-nearest neighbor classification

Classify a point using the labels of its k-nearest neighbors among the training points.

MNIST:

K	1	3	5	7	9	11
Test error (%)	3.09	2.94	3.13	3.10	3.43	3.34

In real life, there's no test set. How to decide which k is best?

❶ Hold-out set:

- Let S be the training set.
- Choose a subset $V \subset S$ as a validation set.
- What fraction of V is misclassified by the k-nearest neighbors in $S \setminus V$?

K-nearest neighbor classification

Classify a point using the labels of its k-nearest neighbors among the training points.

MNIST:

K	1	3	5	7	9	11
Test error (%)	3.09	2.94	3.13	3.10	3.43	3.34

In real life, there's no test set. How to decide which k is best?

❶ Hold-out set:

- Let S be the training set.
- Choose a subset V ⊂ S as a *validation* set.
- What fraction of V is misclassified by the *k*-nearest neighbors in S \ V ?

❷ Leave-one-out cross-validation:

- For each point x ∈ S , find the *k*-nearest neighbors in S \ *{x}*.
- What fraction are misclassified?

Cross-validation

How to estimate the error of *k*-NN for a particular *k*?

10-fold cross-validation

- Divide the training set into 10 equal pieces.
- Training set (call it S): 60,000 points
- Call the pieces S_1, S_2, …, S_{10}: 6,000 points each.

- For each piece S_i:
  - ➢ Classify each point in S_i using *k*-NN with training set $S - S_i$
  - ➢ Let $\in_i$ = fraction of S_i that is incorrectly classified

- Take the average of these 10 numbers:

$$\text{estimated error with } k\text{-NN} = \frac{\in_1 + \cdots + \in_{10}}{10}$$