Machine Learning
DA220

Soumitra Samanta
soumitra.samanta@gm.rkmvu.ac.in
Office: PB405

Classification

Recap

- What is ML?
- Some applications of ML
- Deductive vs Inductive inference
- Different types of data and their representation
- Data similarity: different types of distance metrics

K-nearest neighbour classifier

- Let $X_1, X_2, \ldots, X_n$ be given feature (observation) vectors, $X_i \in R^d$
- Let Y_i denote the class Lebel of X_i
- Let the number of classes be C
  - $Y_i \in \{1,2,...,C\}$
- Let Y_i's are known $\forall i = 1,2,...,n$
- Let X be a vector for which we don't know its class Lebel

k-nearest neighbour classifier (cont.)

- Let k be a (+)ve integer
- Find k- nearest neighbour of X among $X_1, X_2, \ldots, X_n$
- Let k_i of these nearest neighbours belong to i^{th} class for each $i = 1,2,...,C$
  - $\sum_{i=1}^{C} k_i = k$
- Put X in the i^{th}class if $k_i > k_j, \forall i \neq j$

K-nearest neighbour classifier (cont.)

- Remark
  - When $k = 1$, the rule is known as nearest neighbour classifies
  - There is no universally acceptable way of choosing the value of k
  - The value of k depends on data point dispersion not only depend on the number of data points
  - For two different values of k, we may get different results

Classifier (model) evaluation

- Data partition
  - Training
  - Validation
  - Testing
- Model error/loss [$\bar{Y}_i := f(X_i)$]:
  - $\ell(X_i, Y_i, \bar{Y}_i) := \begin{cases} 0 & if\ \bar{Y}_i = Y_i \\ 1 & otherwise \end{cases}$
  - $E\left[f(X, Y)\right] = \frac{1}{n}\sum_{i=1}^{n} \ell(X_i, Y_i, \bar{Y}_i)$
    - Pointwise $0-1$-loss

Recap

- What is ML?
- Some applications of ML
- Deductive vs Inductive inference
- Different types of data and their representation
- Data similarity: different types of distance metrics
- kNN rule classifier
- Classifier evaluation
  - data partition

Assignment-1

- Implement kNN classifier and test on MNIST digit data
  - Download the dataset from here: http://yann.lecun.com/exdb/mnist/
    - Strictly follow their data partition
      - There is no validation set!
      - Make your own validation set from the training set (20%)
  - Use different similarity metrics ($p = 1,\ 2,\ \infty$) and ($k = 1,\ 3, \cdots,\ 25$) calculate the classifier errors
  - Plot (3-D) the classification errors/accuraccy for different p's and k's
- Submission deadline: 21-02-2023