1 Module 1: Intro to ML

Contents

1.1 The Black Box Model

- Treat a problem like a 'black box' →
  - input x (bolded if a list of values, e.g. vector, array, tuple)
  - scalar output $f(x) = y$
  - model attempts to predict values $\hat{y}$
- We evaluate the efficacy of that model with a loss fcn. $L = (y - \hat{y})^2$

1.2 Distribution Random Variables

- Variables in a dataset may be discrete (e.g. coin flip) or continuous (e.g. PDF of height)
- Random variables: sampled from a PDF
  - Variable: Y
  - Individual samples: y
  - Variable distribution: $Y \propto p_y$
- Square brackets denote the size of the sample, e.g. $[y]_5 = [y_1, y_2, ..., y_9] \rightarrow 10$ samples from Y
- Nondeterministic system: outputs are different for same input

1.3 Expected value and variance

- Distributions:
  - Discrete: $p_y(y) \in [0,1]$ and $\sum_i^{\pm\infty} p_{y_i}(y) = 1$
  - Continuous: $p_y(y) \geq 0$ as long as $\int_{\pm\infty} p_y(y)dy = 1$
- For a random var Y distributed as p_y:
  - Expected value (mean) $E[Y] = \mu_y = \sum_i^{\pm\infty} y_i p_{y_i}(y)$ **disc.** $= \int_{\pm\infty} y p_y(y)dy$ **cont.**
  - Variance $Var[Y] = \sigma^2 = E[(Y-\mu_y)^2] = E[Y^2] - (E[Y])^2$
- Law of large numbers $\lim_{n\to\infty} \text{avg.}([y]_n) = E[Y]$

1.4 Intro to pandas

- List homogeneity: all values in that object are the same datatype (e.g. np.arrays are homogeneous) → each col. in a DataFrame is homogeneous
- Columns: a random variable(s) or **features**
- Rows: samples from a distribution
- .read_csv() works on urls!
- DataFrame summaries:
  - .info(): number of entries, cols, datatypes of each col, size
  - .describe(): summarizes numerical data - number of data points, mean, std, quartiles, max/min
  - .head()/ .tail(): top/bottom of DataFrame

1.5 Data selection in pandas

- Bracket selection: table[]
  - Columns: label:str, labels:(list of str)
  - Rows: slice (e.g. table[0:5]) or boolean mask (e.g. table['height'] $\leq$ 6])
- Label-based selection: table.loc[rowselector, columnselector:optional]

  - rowselector: index (int, str, date, etc.), indices, slice (iff index is integer-based), boolean mask
  - columnselector: label:str, labels:(list of str)

- Integer-based selection: table.iloc[rowselector, columnselector]
  - rowselector, columnselector: int, list of int, slice

1.6 Operations and plots in pandas

- Columns:pd.Series (e.g. Series.sum(), etc.)
- add columns: table.add(another_table) or multiply columns: table.multiply(another_table)
- DataFrame.plot(kind(hist, scatter, line, etc.):str, x:str(optional), y:str(optional))