CS4414 Cheat-Sheet

Gurpreet Singh
`gsingh95@uwo.ca`
6. November 2017

Data Preparation

Data Cleaning

Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incorrect parts of the data and then deleting the dirty or coarse data.

Statistics

Definitions

`BoxPlot`: Extended lines mean Min and Max. Box is third quartile, median and first quartile top to bottom. Suspected outliers are empty dots, outliers are solid dots

`Histogram`: Same as a barchart but groups numbers into ranges on the x-axis

`Independence`: Two random variables X and Y are independent if Any conditional distribution of X is the same as the marginal distribution of X and knowing about Y provides no information about X.

`Marginal Distribution`: gives the probabilities of various values of the variables in the subset without reference to the values of the other variables. The sum of the columns.

`Conditional Distribution`: shows the probability that a randomly selected item in a sub-population has a characteristic you're interested in. Moving the column (scanning)

`Joint Distribution`: X and Y have a joint distribution if their realizations come together as a pair. (X, Y) is a random vector, and realizations may be written $(x_1, y_1), (x_2, y_2), ...$, or $\langle x_1, y_1\rangle, \langle x_2, y_2\rangle, ...$

`Sample Mean`: Given a dataset (collection of realizations) $x_1, x_2, ..., x_n$ of X, the sample mean is:

$$\bar{x}_n = \frac{1}{n}\sum_i x_i$$

Given a dataset, $\bar{x}_n$ is a fixed number. It is usually a good estimate of the expected value of a random variable X with an unknown distribution.

`Sample space` $\mathcal{S}$ is the set of all possible events we might observe. Depends on context.
- Coin flips: $\mathcal{S} = \{h, t\}$
- Eruption times: $\mathcal{S} = \mathbb{R}^{\geq 0}$
- (Eruption times, Eruption waits): $\mathcal{S} = \mathbb{R}^{\geq 0} \times \mathbb{R}^{\geq 0}$

`An event` is a subset of the sample space.
- Observe heads: $\{h\}$
- Observe eruption for 2 minutes: $\{2.0\}$
- Observe eruption with length between 1 and 2 minutes and wait between 50 and 70 minutes: $[1, 2] \times [50, 70]$.

Random Variables

`Random variable` is a mapping from the event space to a number (or vector)

`Discrete random variable` take values from a countable set

`Continuous random variable` take values in intervals of $\mathbb{R}$
For a continuous r.v. X, $\Pr(X = x) = 0$ for all x. There is no probability mass function.

Probability Mass Function

For a discrete X, $p_X(x)$ gives $\Pr(X = x)$.
Requirement: $\sum_{x\in\mathcal{X}} p_X(x) = 1$.
Note that the sum can have an infinite number of terms.
- Only works on discrete values
- Sum of probabilities in set needs to be 1

Cumulative Distribution function

For a discrete X, $P_X(x)$ gives $\Pr(X \leq x)$.
Requirements:
- P is nondecreasing
- $\sup_{x\in\mathcal{X}} P_X(x) = 1$

Note: - $P_X(b) = \sum_{x\leq b} p_X(x)$
- $\Pr(a < X \leq b) = P_X(b) - P_X(a)$

Probability Density Function

For continuous X, $\Pr(X = x) = 0$ and PMF does not exist. However, we define the Probability Density Function f_X:
- $\Pr(a \leq X \leq b) = \int_a^b f_X(x)\,\mathrm{d}x$

Requirement:
- $\forall x\ f_X(x) > 0$, $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$
- Requires a range, doesnt work with discrete values

Cumulative Distribution Function

For a continuous X, $F_X(x)$ gives $\Pr(X \leq x) = \Pr(X \in (-\infty, x])$.
Requirements:
- F is nondecreasing
- $\sup_{x\in\mathcal{X}} F_X(x) = 1$

Note:
- $F_X(x) = \int_{-\infty}^{x} f_X(x)\,\mathrm{d}x$
- $\Pr(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$

Supervised Learning

Definitions

`Supervised Learning` is when correct answers are given to a training model and it uses them to make future perdictions following similar patterns

`Columns` are called input variables or features or attributes

`Labels` or output variables are the outcome we are trying to perdict

`Training Example` or instance is a row in a table

`Data Set` is the whole table

h is called a predictive model or `hypothesis`

`Classification` Is this A, B or C? (Binary if only 2 categories)
`Regression` How much or how many?
`Clustering` How is this organized?
`Reinforcement` What should i do next?

Example Problem

Given a data set $D \subset (\mathcal{X} \times \mathcal{Y})^n$, find a function:

$$h : \mathcal{X} \to \mathcal{Y}$$

such that $h(\mathbf{x})$ is a good predictor for the value of y.

Steps to solve a supervised learning problem

1. Decide what the input-output pairs are.
2. Decide how to encode inputs and outputs. This defines the input space $\mathcal{X}$, and the output space $\mathcal{Y}$.
3. Choose model space/hypothesis class $\mathcal{H}$.
4. Choose an error function (cost function) to define the best model in the class
5. Choose an algorithm for searching efficiently through the space of models to find the best.

Linear Hypothesis

Suppose y was a linear function of $\mathbf{x}$:

$$h_{\mathbf{w}}(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + \cdots$$

w_i are called parameters or weights. Typically include an attribute $x_0 = 1$ (also called bias term or intercept term)

Choosing Weights

Error Minimization

Define an error function or a cost function to measure how much our prediction differs from the 'true' answer on the training data. The weights should make the prediction as close to the true values as possible.

Picking the error function

Least Means Squares

Draw a linear line such that the sum of the distance between each the line and each training point (error) is as small as possible.

Linear Regression Summary

The optimal solution (minimizing sum-squared-error) can be computed in polynomial time in the size of the data set.
A very rare case in which an analytical, exact solution is possible

Improving Linear Regression

1. Explicitly transform the data, i.e. create additional features
- Add cross-terms, higher-order terms
- More generally, apply a transformation of the inputs from $\mathcal{X}$ to some other space $\mathcal{X}'$, then do linear regression in the transformed space
2. Use a different model space/hypothesis class