ML and Numerical Software Development

Statistics-I

Organon Analytics

November 19, 2019

Statistics Agenda

- Data generating processes and some notation
- Types of (structured) data in Statistics: Categorical, Ordinal, Scale
- The subjects of Statistics. Descriptive Statistics and Inferential Statistics
- Univariate statistics: Mean, variance, quantiles and all that. Multi-modality, outliers, missing values
- Bivariate statistics: Measures of dependency (Linear and non-linear measures). 3 cases: categorical-categorical, categorical-scale, scale-scale
- The central problem of Statistics and ML: Estimation from finite data. Estimation methods in statistics
- Maximum Likelihood Estimation: The *go-to* estimation method in Machine Learning. A deep dive: formulation, examples

Statistics Agenda

- The measures of the distance (discrepancy) between two distributions (i.e. two random variables). Entropy, cross-entropy. Use cases in ML
- Statistical dependency, correlation and causation. Definitions, pitfalls, examples
- Hypothesis Testing, Type-I, Type-II errors
- Multivariate Linear Regression: The grandpa of all (supervised) ML algorithms: Specification, estimation, hypothesis testing, feature selection, measuring, goodness-of-fit, prediction
- Ugly facts and how to handle them: Outliers, missing values, categorical variables, non-linearities

Data generating processes and some notation

- Statistics and ML are interested in data generating processes (DGPs)
- DGPs are everywhere: Physics, Economics, Finance, Marketing, etc. We are drowned in the data
- A **sample** is a collection of data points generated by a DGP. It is important that this data be **representative** of all data that -potentially- **could** be produced by the DGP
- Mathematically, a single data point is represented as a tuple-of-values, i.e. a vector:

$$\mathbf{x} \equiv \{x_1, x_2, \cdots, x_d\}$$

  where d denotes the dimension of the data point.
- A sample $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$ might consist of many of these data points: N is the standard letter to denote the **sample size**
- N is important: The knowledge that can be extracted from the data is proportional to N

Data generating processes and some notation

- We reserve the bold-capital-letter X for the vector of random variables:

$$\mathbf{X} \equiv \{X_1, X_2, \cdots, X_d\}$$

- X_i is a random variable; x_i is a value (a realization of, an instance of) of this r.v. X is one; x is many.
- Some X_i are input(s), some are outputs. Output RVs are usually denoted by $\mathbf{Y}$. Output variables are also called **dependent** variables, and input variables are also called **independent** variables. Note that there might not be any output at all.
- Hence, the sample data could be denoted by

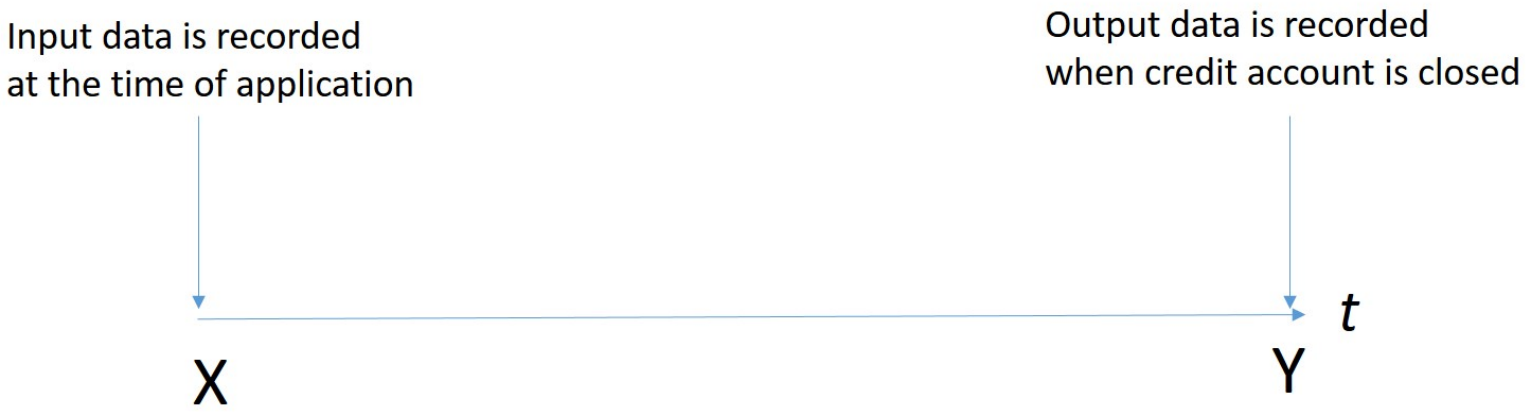
$$\text{Data} = (\mathbf{x_i}, \mathbf{y_i}), i = 1, 2, \cdots, N$$

Data generating processes and some notation

- Remember: Joint distribution(or density) function defines the DGP: That is all we need.
- But we do not know the joint CDF or joint PDF
- All we have is some finite data, which are generated by an unknown joint CDF(PDF)
- Statistics is -mostly concerned- understanding *features of a DGP* by the data it produces
- *Features*: The entire distribution, its central value, its dispersion, etc.

Data generating processes: Examples

- Credit Risk Application process:
  - **X**: Age, profession, occupation, marital status, gender, financial data, data on past credit applications and their performances, social network data, etc.
  - **Y**: Defaulted on the account (Yes or No)



- Taking a picture:
  - **Grayscale**: Produces a vector of HW random variables where H, W corresponds to height and width of the picture and determines by the resolution. Each variable takes integer values in the interval $[0, 256]$
  - **Colored**: Produces a vector of $3HW$ random variables

Data generating processes: Examples

- Surfing an e-commerce cite:
  - **X**: Membership data, clickstream data (which page, when, duration, mouse behavior), referral page
  - **Y**: Add the item to shopping cart, purchase item, purchase amount of the purchased item
- Manufacturing:
  - **X**: Sensor data on manufacturing line (temperature, humidity, vibration, camera data, etc.), parts data (type, date of production, last repair date, etc.)
  - **Y**: Malfunction in one part
- Healthcare:
  - **X**: Patients past history (diagnoses, treatments, drugs, analyses, MR, X-Ray data, etc.), relatives past history
  - **Y**: Diagnosed disease, death

Data Types: { Categorical, Ordinal, Scale }

Categorical data:

1. Discrete
2. No natural ordering exists between its values

Examples:

- Gender: {M, F, U, X, ?} (Categories {U, X, ?} are the result of recording process)
- Marital Status: {Single, Married, Divorced, Unknown}
- Residential city: {Istanbul, Ankara, ...}
- Zip-codes: {Istanbul, Ankara, ...}

Categorical data is represented by discrete random variables!

Data Types: { Categorical, Ordinal, Scale }

Ordinal data:

1. Discrete
2. A natural ordering exists between its values, BUT
3. No proper distance function could be defined between its values

Examples:

- Rating: {Dislike, Neutral, Like} (Dislike $<$ Neutral $<$ Like)
- School: {None, Primary, Secondary, High-school, College, University}

Ordinal data is represented by discrete random variables!

Data Types: { Categorical, Ordinal, Scale }

Scale data:

1. Takes values on a subset of R^1 with "potentially" infinite values
2. A natural ordering exists between its values
3. A distance function exists between its values
4. Two sub-types: **Interval** (No natural zero) and **Ratio** (A natural zero exists)

Scale data is represented by continuous random variables!

Expect all data types in a real data-set

Data Types: Why do we care?

1. The available statistics are different for different types.

Statistics	Categorical	Ordinal	Scale
Frequencies (mode, histogram)	✓	✓	✓
Order statistics (median, quantiles, etc.)	X	✓	✓
Moments (mean, variance, etc.)	X	X	✓

2. If it exists, the type of output variable determines the type of algorithm that will be used to model the dependencies between the output and inputs

Output type	Algorithms
Categorical	Logistic Regression, Multinomial Regression
Ordinal	Probit, Quantile Regression
Scale	Linear Regression

More on this later on!

Statistics: What it is about?

- **Sample&Collect**: Design a representative sample and collect data (Hardest part)
- **Organize**: Design data model for access
- **Data Quality**: Detect& Correct errorenous data
- **Visualization**: Visualize the data and the results
- **Analysis**: Apply Descriptive Statistics and Inferential Statistics algorithms
- **Interpret and Communicate**: Interpret results and communicate

Statistics: Descriptive and Inferential Statistics

1. **Descriptive Statistics**: Summarize the data. Purposes:
   - To gain a 1-d (univariate) and 2-d (bivariate) understanding of the data
   - To detect data quality errors and correct them
   - To transform the variables for better inferential statistics
   - Useful when presenting and communicating: a) Hard to think in multi-dimensions, b) Hard to reason about a distribution
2. **Inferential Statistics**: Use **data** to arrive at the **model** (hence DGP). Use model to **understand**, **predict** and **simulate**
   - To understand complex dependencies in multivariate data
   - To predict for the new inputs
   - To simulate the system

Descriptive Statistics: Important questions

- Is the distribution nice? (Nice: uni-modal, no-outliers, symmetric, light tails)
- Are there missing values? Density of missing values?
- What is the cardinality of a discrete variable?
- Is data asymmetrical?
- Does data have outliers?
- Is there multi-modality in the data?
- Are two variables correlated? What is the degree of correlation?
- Are two variables independent? What is the degree of independence?

Univariate Statistics: Centrality, dispersion, asymmetry, outliers

- Centrality: Mean, sample mean

$$\begin{aligned} \mu &= E[X] = \int x f_X(x) dx \\ \hat{\mu} &= \bar{X} = \frac{1}{N} \sum_{i=1}^{n} x_i \end{aligned}$$

- Dispersion: Variance, sample variance:

$$\begin{aligned} \sigma^2 &= E[(X-\mu)^2] = \int (x-\mu)^2 f_X(x) dx \\ \hat{\sigma^2} &= S_n^2 = \frac{1}{N-1} \sum_{i=1}^{n} (x_i - \bar{X})^2 \end{aligned}$$

- Asymmetry: Skewness, sample skewness:

$$\kappa = \frac{E[(X-\mu)^3]}{\sigma^3} = \frac{\int (x-\mu)^3 f_X(x)dx}{\sigma^3}$$

$$\hat{\kappa} = \frac{\frac{1}{N}\sum_{i=1}^{n}(x_i-\bar{X})^3}{\hat{\sigma}^3}$$

  Negative skewness: Right leaning pdf (for unimodals)
  Positive skewness: Left leaning pdf

- Tail statistics: Kurtosis, sample kurtosis

$$\kappa = \frac{E[(X-\mu)^4]}{\sigma^4} = \frac{\int (x-\mu)^4 f_X(x)dx}{\sigma^4}$$

$$\hat{\kappa} = \frac{\frac{1}{N}\sum_{i=1}^{n}(x_i-\bar{X})^4}{\hat{\sigma}^4}$$

  A value greater than 3 implies fat tails (large values on both ends)

Univariate Statistics: Order statistics

Defined in terms of CDF function $F_X(x)$

- Minimum $= F^{-1}(-\infty)$
- Maximum $= F^{-1}(\infty)$
- Median $= F^{-1}(0.5)$
- 1st quartile $= F^{-1}(0.25)$
- 3rd quartile $= F^{-1}(0.75)$
- Inter-quartile range (IQR) $= F^{-1}(0.75)$ - $F^{-1}(0.25)$
- General p-quantile: $Q_p = F^{-1}(p)$

Univariate Statistics: Recipe for categorical&ordinal data

1. Extract the histogram(Probability mass function)
2. Cardinality:
   (i) Low: If sparse categories exist, merge them. If not do not do anything. If used for regression purposes convert it to other variables using One-hot-encoding
   (ii) High(More than 10 categories): Most problematic from a modelling point-of-view. Merge sparse categories under a common category. If used in a regression context, use one-hot-encoding or weight-of-evidence transformations
3. Watch for multiple missing values: Merge them if they do not represent different reasons
4. Note-1: Merging for ordinal data should only be done for consecutive categories
5. Note-2: If an output variable Y exists, a supervised-merging of the categories is possible (Optimal binning)

Univariate Statistics: Recipe for scale data

1. Extract the binned-histogram(discretized pdf)
2. Visual examination of histogram: Reveals asymmetry, heavy outliers
3. Compute inter-quartile range $IQR = Q_{0.75} - Q_{0.25}$. Confidence interval :

$$[Q_{0.25} - 1.5 * IQR, Q_{0.75} + 1.5 * IQR]$$

   Points outside this interval are outliers
4. Handling outliers
   - (i) Statistics such as mean, variance are not reliable. Use robust-statistics such as trimmed mean
   - (ii) Regression: If the variable is output variable a) Delete the samples if the number is small, b)Replace with a fixed value if number is large (do not lose sample)
   - (iii) Regression: If the variable is input variable, bin the variables by using a recursive binning algorithm

Univariate Statistics: Missing values

1. Extract the distribution of missing values. Merge them under a single category if they do not corresponds to different reasons
2. If the density of missing values is very high(more than 95%), question the data source and validity of the variable: You might skip it for further analysis
3. If used in a regression context:. Replace the missing value:
   - *i* With central value: Replace with the central value (sample mean of non-missing values)
   - *ii* By imputation(Advanced): Fit a model by using the variable as output
   - *iii* Weight-of-evidence variable: Use $E[Y|X = x]$ instead of X. You can compute $E[Y|X = x]$ by any univariate smoother
   - *iv* Create two variables: a) A scale variable of non-missing values b) A scale variable with two values (1 if missing; 0 if not-missing)

Bivariate Statistics: Measures of dependency between two variables

Scale-scale

- Pearson correlation coefficient: A measure of linear dependency. Not robust. Sensitive to outliers

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Sample estimate is given by

$$\hat{\rho}_{XY} = \frac{\sum_i (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_i (x_i - \bar{X})^2} \sqrt{\sum_i (y_i - \bar{Y})^2}}$$

- $\hat{\rho}_{XY} \sim 0$: weak linear dependency. Check for non-linear dependency
- $\hat{\rho}_{XY}$ large: strong linear dependency
- You should check for non-linear dependency in any case

Bivariate Statistics: Measures of dependency between two variables

Scale-scale (Advanced)

- Always check for non-linear dependency
- **Method-1**: (X, Y). Use Regression Tree. Use the resulting goodness-of-fit (R-squared value) as a measure of dependency between the two variates
- **Method-2**: Use cubic-spline interpolation:

$$Y = \beta_0 + \sum_i \beta_k B_k(x)$$

Use the resulting goodness-of-fit (R-squared value) as a measure of dependency between the two variates

Bivariate Statistics: Measures of dependency between two variables

Categorical-categorical

$$\begin{aligned} p_{ij} &= P(X_i = x_i, Y_j = y_j) \text{ (joint probabilities)} \\ p_i &= P(X_i = x_i) \text{ marginal probs for } X \\ p_j &= P(Y_j = y_j) \text{ marginal probs for } Y \end{aligned}$$

- Kendall's tau:

$$\tau = \frac{\sum_i \sum_j p_{ij}^2 / p_i - \sum_j p_j^2}{1 - \sum_j p_j^2}$$

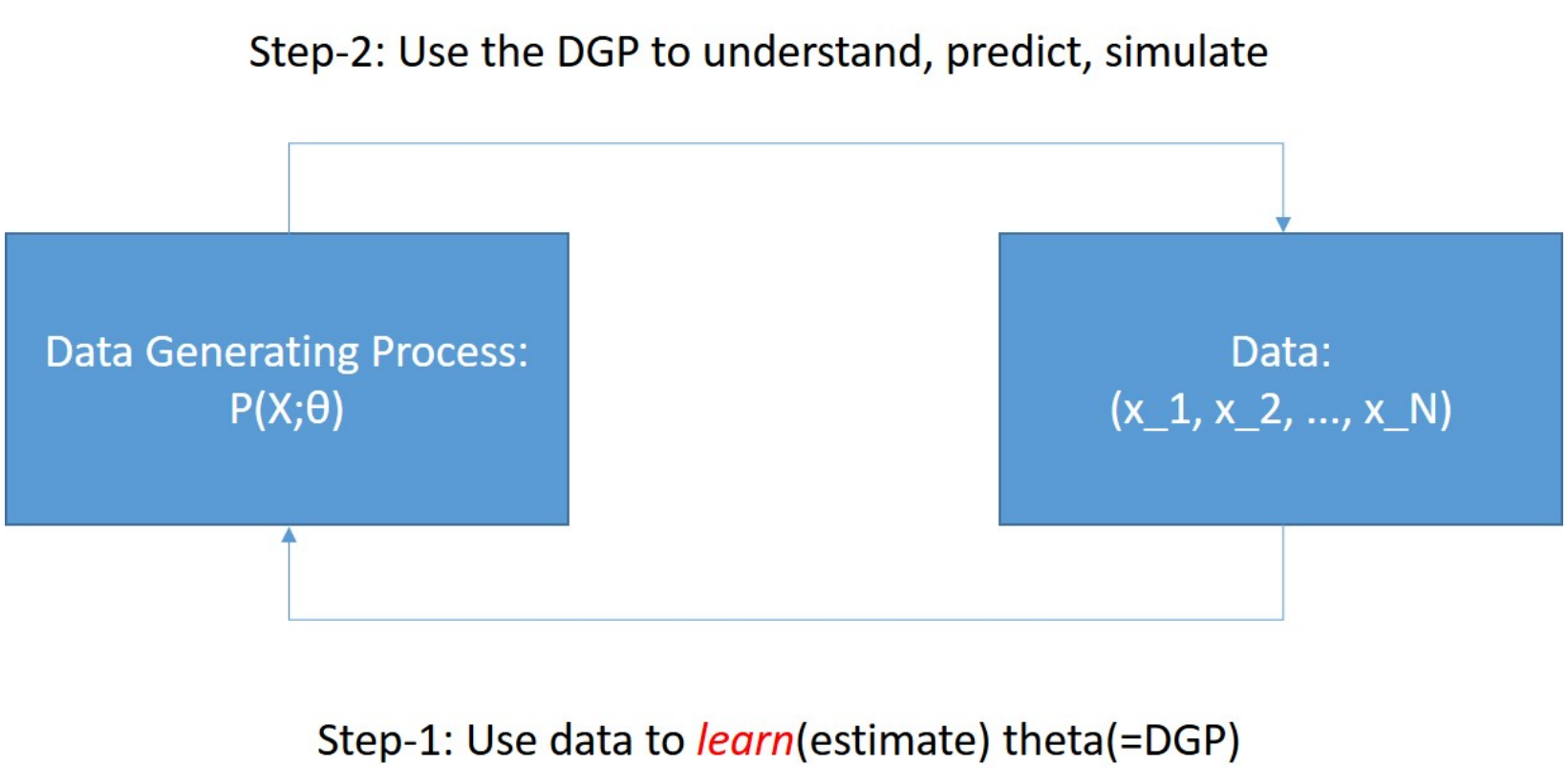
  $\tau = 0$ is independence, $\tau = 1$ is 1-to-1 correspondence between variable categories
- Uncertainty coefficient

$$U = \frac{-\sum_i \sum_j p_{ij} \log(p_{ij}/p_i p_j)}{\sum_j p_j \log(p_j)}$$

Bivariate Statistics: Measures of dependency between two variables

Output: Binary Categorical
Input: Scale
IMPORTANT CASE. ROC-statistics is a non-parametric measure.
Will be defined after Type-I and Type-II errors are defined.

Statistical Inference: Use data to **infer** DGP



The parameter vector θ completely specifies the DGP.

(Parametric) Statistical Inference: Find a *good* estimator $\hat{\theta}$ for the unknown parameters θ

Statistical Inference: The form of $p(X;\theta)$ is known

There are good estimators that will give you

$$\lim_{N\to\infty} \hat{\theta} = \theta$$

Hence the error will get monotonically smaller with more data
Example: Data is known to be generated by a normal distribution. Estimate μ, σ^2) by assuming that $p(x;\theta)$. Nice and easy!

Statistical Inference: The form of $p(X;\theta)$ is unknown

You will have two kind of errors:

(i) Specification error: Since you do not know the exact mathematical form of $p(X;\theta)$

(ii) Sampling error: Since you have finite data.

Example: Data is generated by an exponential distribution but you don't know it and try to **fit** a normal distribution. The parameters you found $(\hat{\mu}, \hat{\sigma^2})$ do not produce the data.
Try the following:

1. Try different known specifications
2. Try to fit a mixture distribution if(1) fails
3. Try a non-parametric specification if (2) fails

Statistical Inference

Given a sample $X_1, X_2, \cdots, X_N$

1. Infer population parameters of $p(X)$
2. Infer a functional of $p(X)$ (e.g. mean, variance, etc.)

1. (One-dimensional Parametric Estimation): Let $X_1, X_2, \cdots, X_N$ be independent Bernoulli(p) observations. The problem is to estimate the parameter p
2. (Multi-dimensional Parametric Estimation): Let $X_1, X_2, \cdots, X_N$ be independent $N(\mu, \sigma^2)$. The problem is to estimate the parameter (μ, σ^2)

Statistical Inference

3. (Non-parametric Estimation): Let $X_1, X_2, \cdots, X_N$ be the sample. The problem is to estimate the cumulative distribution function $F_X(x) = (X <= x)$
4. (Non-parametric density estimation): Let $X_1, X_2, \cdots, X_N$ be the sample. The problem is to estimate the probability density function $f_X(x)$

5 Regression: Assume data is $(X_1, Y_1), (X_2, Y_2), \cdots, (X_N, Y_N)$ where X is called input (also predictor, regressor, independent) variable, and Y is called output (also outcome, response, or dependent) variable.
Estimate $f(X) = E[Y|X = x]$

5.1.1 Linear Regression(Parametric)

$$f(X) = \beta_0 + \sum_i \beta_i X_i$$

5.1.2 Generalized Additive Modelling(Parametric):

$$f(X) = \beta_0 + \sum_{ij} \beta_{ij} f_{ij}(x_i)$$

5.1.3 Regression Trees(Non-parametric):

$$f(X) = \sum_i \beta I_A(X)$$

6 Classification:
Estimate
$f_k(X) = P(Y = C_k|X = x)$ for each class $C_k, k = 1, \cdots, M$

i Binary Logistic Regression(Parametric)

$$f(X) = P(Y = 1|X = x) = \frac{1}{1 + e^{-(\beta_0 + \sum_i \beta_i X_i)}}$$

where $Y \in \{0, 1\}$

ii Multivariate Logistic Regression(Parametric)

$$f_k(X) = P(Y = C_k|X = x) = \frac{1}{1 + e^{-(\beta_0 + \sum_i \beta_{ik} X_i)}}$$

where $Y \in \{C_1, C_2, \cdots, C_M\}$

iii Artifical Neural Networks (M-layer)

$$f_k(X) = P(Y = C_k|X = x) = \frac{1}{1 + e^{-Z_{kM}(X)}}$$

where $Z_{kM}(X)$ is defined recursively

Estimators: Functionals of PDF

A statistical functional $T_X(f)$ is a functional of the CDF F of random variable X with the following from

$$T(F) = E[r(X)] = \int r(x) dF_X(x) dx$$

Use the plug-in estimator

$$T(\hat{F}_N) = \frac{1}{N} \sum_i r(X_i)$$

Examples:

1. Mean: $r(x) = X$

$$\bar{X} = \frac{1}{N} \sum_i X_i$$

2. Variance: $r(x) = (X - E[X])^2$

Estimators: Regression problems (Y is scale variable)

Assume additive randomness:

$$\begin{aligned} Y &= f(X) + \epsilon \\ E[\epsilon|X] &= 0 \end{aligned}$$

Least Squares Estimator:

$$\hat{f}_{LSE}(x) = \arg\min_{f} E[Y - f(X)^2]$$

Minimizing the above functional, one gets:

$$\hat{f}_{LSE}(x) = E[Y|X = x]$$