Statistical Methods for Data Science: A Starter Kit

Yinan Yu

yinan@chalmers.se/yinan.yu@asymptotic.ai

Statistical Data Type (l1)

Categorical data: labels or tags

- Nominal: unordered labels, e.g. species of ducks
- Ordinal: ordered labels, e.g. {“duckling”, “teen duck”, “adult duck”}

Numerical data:

- Discrete (interval): countable, e.g. integers; numbers of ducks
- Continuous (ratio): uncountable, e.g. real values; weights of ducks

Data Container (l1)

Array (tensor):

- Scalar, vector, matrix, higher order array
- Same numerical data type
- Python library: numpy

Table:

- Described by columns and rows
- Mixed data types
- Python library: pandas

Descriptive Statistics: numerical data (l1)

Data set (a sample): numerical data $x_1, \cdots, x_N$

Centrality:

- sample mean: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- median: sort x_i and median is the value in the middle
- mode (discrete values): the most frequent value in a sample

Dispersion:

- min, max, range: $\min\{x_i\}$, $\max\{x_i\}$, $\max\{x_i\} - \min\{x_i\}$
- quantiles/percentiles: given $p \in (0,1)$, q is a p-quantile of the data if $p \times 100\%$ of the data are smaller than q
- sample variance: $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$
- sample standard deviation: s

Dependence: given a data set with two paired values:

$$\{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$$

- covariance:

$$cov(x, y) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

- correlation: measures how close data is to a linear relationship

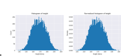
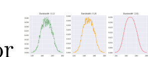
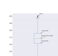
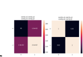
$$corr(x, y) = \frac{cov(x, y)}{\sigma_x \sigma_y}, \quad -1 \le corr(x, y) \le 1$$

Descriptive Statistics: categorical data (l1)

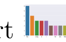
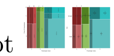
Data set (a sample): categorical data $x_1, \cdots, x_N$

- Count/frequency
- Transformed into numerical, discrete data

Visualization: numerical data (l1)

- Distribution:
  - Histogram/normalized histogram 
  - Kernel density estimator 
  - Box plot 
- Dependence (two variables):
  - Scatter plot 
  - Heat map for covariance/correlation 

Visualization: categorical data (l1)

- Distribution
  - Bar chart 
  - Pie chart 
- Dependence
  - Mosaic plot 

Probability distribution (l2)

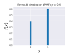
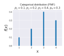
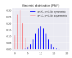
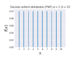
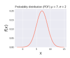
- Experiment: an action that leads to one outcome
- Sample space: the set of all possible outcomes from an experiment
- Event: a subset of the sample space
- Random variable (discrete/continuous): assigning a numerical value to each outcome of the experiment; denoted by capital letters, e.g. X
- Probability distribution: the probability of the occurrence of *any* event in the sample space; can be described by P(event)/PDF/PMF/CDF
  - P(event): the probability of an event occurring
  - PDF $f(x)$: the probability density function for continuous random variables; $\int_{-\infty}^{+\infty} f(x)dx = 1$
  - PMF $f(x)$: the probability mass function for discrete random variables; $\sum_{x=-\infty}^{+\infty} f(x) = 1$
  - CDF $F(x)$: the cumulative density function; $F(x) = P(X \le x)$
- Quantile function Q: the inverse CDF, i.e.

$$F_X(Q(p)) = p \text{ and } Q(F_X(q)) = q$$

- Conditional probability
- Independent and identically distributed (i.i.d.) random variables

Examples (l2)

Discrete/continuous, PMF/PDF, parameters, typical use cases (statistical data type, example scenarios)

- Bernoulli distribution 
- Categorical distribution 
- Binomial distribution (l8) 
- Discrete uniform 
- Gaussian distribution 

Generalize this learning routine to unknown distributions

Properties of Gaussian distributions (l6)

- Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ be a Gaussian random variable, then the following random variables are also Gaussian
  - Scaling: $tX \sim \mathcal{N}(t\mu_X, t^2\sigma_X^2)$, $t \neq 0$ is a constant
  - Translation: $X + c \sim \mathcal{N}(\mu_X + c, \sigma_X^2)$, c is a constant
  - $tX + c \sim \mathcal{N}(t\mu_X + c, t^2\sigma_X^2)$

- Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ be two **independent** Gaussian random variables, then the following random variables are also Gaussian
  - $X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$
  - $X - Y \sim \mathcal{N}(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$

Bayes' rule (l4, l5)

- Parameter estimation:

$$f_{\Theta|data}(\theta \mid data) = \frac{\overbrace{f_{data|\Theta}(data \mid \theta)}^{\text{likelihood}}\ \overbrace{f_\Theta(\theta)}^{\text{prior}}}{f_{data}(data)}$$

  where $f(\cdot)$ is the PDF/PMF

- Multinomial naive Bayes classifier:

$$P(Y = y \mid X = x) = \frac{\overbrace{P(X = x \mid Y = y)}^{\text{likelihood}}\ \overbrace{P(Y = y)}^{\text{prior}}}{P(X = x)}$$

- Gaussian naive Bayes classifier/Gaussian mixture models:

$$P(Y = y \mid X = x) = \frac{\overbrace{f_{X|Y=y}(x \mid Y = y)}^{\text{likelihood}}\ \overbrace{P(Y = y)}^{\text{prior}}}{f_X(x)}$$

Q-Q plot (l3)

- Use cases:
  - Compare a data distribution to a theoretical distribution (one sample test)
  - Compare two data distributions (two sample test)
- Steps:
  - Choose a set of m probabilities $p_1, \cdots, p_m \in [0, 1]$ (make sure they spread evenly between 0 and 1)
  - For $i = 1, 2, \cdots, m$:
    * Compute the quantile q_i^1 of the first distribution at p_i
    * Compute the quantile q_i^2 of the second distribution at p_i
    * Make a scatter plot of the pair (q_i^1, q_i^2)
- Interpretation
  - Case 1: if the two distributions are identical, the points in the Q-Q plot should follow a 45° straight line $y = x$
  - Case 2: if the two distributions are linearly related, the points in the Q-Q plot follow a straight line that is not necessarily $y = x$
  - Case 3: if the two distributions are from different families of distributions, the points in the Q-Q plot are not lying on a straight line.

Mathematical Modeling (l3)

$$y = g(x; \theta \mid h)$$

1. What do we want to predict, i.e. what is the target y?
2. What are the variables x?
3. What is the mathematical function g that relates variables x to the target y?
4. Are there any hyperparameters h in the function g? How do we choose them?
5. What are the unknown parameters θ in g? **How do we estimate them from data?**

Parameter estimation (l4)

- Maximum likelihood estimation: frequentist approach - **θ is deterministic**
- Maximum A Posteriori estimation: Bayesian approach - **θ is probabilistic**

Maximum Likelihood Estimation (l4)

Given a model $y = g(x; \mathcal{O} \mid h)$, where $\mathcal{O}$ is a set of parameters

a) Describe the experiments

b) Describe the data generated from the experiments

c) Describe the random variables (typically with i.i.d. assumption)

d) Choose a parameter of interest $\theta \in \mathcal{O}$

e) Choose the maximum likelihood estimation as the estimation method:
Given data $x_1, \cdots, x_N$ and assume i.i.d. random variables X_i with PDF/PMF $f(x_i)$,

$$L(\theta \mid x_1, \cdots, x_N) = \prod_{i=1}^{N} f(x_i; \theta)$$

f) Compute $\hat{\theta}_{MLE}$ by maximizing the likelihood function:

$$\begin{aligned} \hat{\theta}_{MLE} &= \arg\max_{\theta} L(\theta \mid x_1, \cdots, x_N) \\ &= \arg\max_{\theta} \prod_{i=1}^{N} f(x_i; \theta) \end{aligned}$$

or equivalently, minimizing the **negative log likelihood function**:

$$\hat{\theta}_{MLE} = \arg\min_{\theta} \; -\sum_{i=1}^{N} \log\left(f(x_i; \theta)\right)$$

- Simple case, e.g. i.i.d. Gaussian, find the closed-form solution by:
  - Taking the partial derivative with respect to the parameter
  - Setting the derivative to zero
  - Solving for the parameter
- In general, the estimate needs to be found by iterative methods, e.g. gradient descent

Maximum A Posteriori Estimation (l4)

Given a model $y = g(x; \mathcal{O} \mid h)$, where $\mathcal{O}$ is a set of parameters

a) Describe the experiments

b) Describe the data generated from the experiments

c) Describe the random variables (typically with i.i.d. assumption)

d) Choose a parameter of interest $\theta \in \mathcal{O}$

e) Choose the maximum a posteriori estimation as the estimation method

- **θ is assumed to be drawn from a random distribution**
- Choose a prior distribution for θ along with the hyperparameters: $f_\Theta(\theta)$
  - Prior might be known by the problem setup
  - If prior unknown, conjugate priors are typically chosen for various reasons
- Find the likelihood function: $f_{X|\Theta}(\boldsymbol{x} \mid \theta)$ (same as in MLE)
- Express the posterior distribution in terms of the prior and the likelihood function

$$f_{\Theta|X}(\theta \mid \boldsymbol{x}) = \frac{f_{X|\Theta}(\boldsymbol{x} \mid \theta) f_\Theta(\theta)}{f_X(\boldsymbol{x})}$$

f) Compute $\hat{\theta}_{MAP}$ by maximizing the posterior function (or equivalently, minimizing the negative log posterior function without the normalization constant). The optimal solution can be found by a closed-form expression or using iterative techniques.

Standardization (l6)

Standardization: let X be a random variable that follows **any probability distribution** with mean μ and standard deviation σ. The standardization of X is

$$Y = \frac{X - \mu}{\sigma}$$

Central limit theorem (l6)

Given an i.i.d. sample $X_1, X_2, \cdots, X_N$ from **ANY probability distribution** with *finite mean μ and variance σ^2* (most distributions satisfy this!), when the sample size N is sufficiently large, the **sample mean** approximately follows a Gaussian distribution with mean μ and variance $\frac{\sigma^2}{N}$, i.e.

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$$

Confidence interval (l6)

- **Data**: $x_1, \cdots, x_N$
- **Random variable**: $X_1, \cdots, X_N$ with i.i.d. assumption
- **Parameter of interest**: θ, e.g. the mean μ
- **Estimate**: $\hat{\theta}$, e.g. the sample mean $\bar{x}$
- **Confidence interval** for a given confidence level $1-\alpha$ (e.g. 95%)
  - Definition:

    confidence interval = ($\hat{\theta}$ - **margin of error**, $\hat{\theta}$ + **margin of error**)

    where

    margin of error = critical value $\times$ standard error of $\hat{\theta}$
  - Calculation:

Distribution of X_i	Scenario	θ	$\hat{\theta}$ (**sampling distribution**)	Critical value	Standard error	Confidence interval	Note
i.i.d. Gaussian	σ known	mean	sample mean $\bar{x}$	$z_{\alpha/2}$	$\frac{\sigma}{\sqrt{N}}$	$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right)$	exact
	σ unknown		(**Gaussian distribution**)	$t_{\alpha/2}$	$\frac{s}{\sqrt{N}}$	$\left(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{N}}, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{N}}\right)$	
i.i.d.	σ known		sample mean $\bar{x}$	$z_{\alpha/2}$	$\frac{\sigma}{\sqrt{N}}$	$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{N}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right)$	approximate
	σ unknown		(**approximately Gaussian under CLT**)	$t_{\alpha/2}$	$\frac{s}{\sqrt{N}}$	$\left(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{N}}, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{N}}\right)$	for large N
i.i.d.	-	any	MLE (**asymptotically Gaussian**)	$z_{\alpha/2}$	$\frac{1}{\sqrt{NI_N(\hat{\theta})}}$	$\left(\hat{\theta} - z_{\alpha/2}\frac{1}{\sqrt{NI_N(\hat{\theta})}}, \hat{\theta} + z_{\alpha/2}\frac{1}{\sqrt{NI_N(\hat{\theta})}}\right)$	asymptotic
i.i.d.	-	any	any statistic (**any distribution**)	bootstrap the error quantile		$\left(\hat{\theta} - \epsilon_{\alpha/2}, \hat{\theta} + \epsilon_{1-\alpha/2}\right)$	approximate

where σ is the standard deviation of the X_i and s the sample standard deviation

Hypothesis testing steps (l7)

- Step 1 Make a "boring" statement
- Step 2 Design an **experiment**
- Step 3 Describe the **data** generated from the experiment and the corresponding random variables
- Step 4 Describe the parameter of interest and their estimates
- Step 5 Translate the "boring" statement into a statistical hypothesis and call it the **null hypothesis** H_0
- Step 6 Find the expression for the **test statistic** s
- Step 7 Find the expression for the **null distribution**
- Step 8 Define **an alternative hypothesis** H_A: one-tailed or two-tailed
- Step 9 Choose a **significance level** α (the tail), which defines the **rejection region**
- Step 10 Collect **data**
- Step 11 Compute the test statistic from data
- Step 12 Compute the p-value
- Step 13 If p-value$< \alpha$, i.e. the test statistic falls in the rejection region of the null distribution, then we reject the hypothesis H_0; otherwise, we fail to reject H_0.

Statistical tests (l8)

Test	Description	Assumption	Test statistic	Null distribution
One-sample z-test	Compare sample mean to a constant; known σ	Large sample or Gaussian	$z = \frac{\bar{x}-c}{\sigma/\sqrt{N}}$	Standard Gaussian
Two-sample z-test	Compare two sample means; known σ(s)	Large samples or Gaussian	$z = \frac{\bar{x}-\bar{y}-c}{\sqrt{\frac{\sigma_X^2}{N_X}+\frac{\sigma_Y^2}{N_Y}}}$	Standard Gaussian
One-sample t-test	Compare sample mean to a constant; unknown σ	Large sample or Gaussian	$t = \frac{\bar{x}-c}{s/\sqrt{N}}$	Student-t
Two-sample t-test	Compare two sample means; unknown σ(s)	Large samples or Gaussian	$t = \frac{\bar{x}-\bar{y}-c}{\sqrt{\frac{s_X^2}{N_X}+\frac{s_Y^2}{N_Y}}}$	Student-t
Paired t-test	Compare sample mean of differences to a constant	Large sample or difference Gaussian	$t = \frac{m_{X-Y}-c}{s_{X-Y}/\sqrt{N}}$	Student-t
Exact Binomial test	Compare estimated success rate $\frac{k}{N}$ to a constant	Small sample	k	Binomial
Approximate Binomial test		Large sample	$z = \frac{k-N\pi}{\sqrt{N\pi(1-\pi)}}$	Standard Gaussian
Exact McNemar's test	Test if an action have different effects on two different groups	Small discordance $n_{01}+n_{10}$	$n_{01}+n_{10}$	Binomial
Approximate McNemar's test		Large discordance $n_{01}+n_{10}$	$\min(n_{01}, n_{10})$	$\mathcal{X}^2$

Machine learning: classification

Multinomial naive Bayes classifier (l5)

- **Prediction y:** categorical data $y \in \{1, \cdots, C\}$
- **Variables x_i, $i = 1, \cdots, n$:** categorical data $x_i \in V$, where V is the vocabulary $V = \{w_1, \cdots, w_K\}$ given K unique categories
  - **Assumptions:**
    * x_i's are independent - **NAIVE!**
    * x_i follows a categorical distribution

  Note: here n is the size of the input data, e.g. the length of a document
- **Model g:**

$$\hat{y} = g(x_1, \cdots, x_n) = \underset{c \in \{1, \cdots, C\}}{\arg\max} P(c) \prod_{i=1}^{n} P(x_i \mid c)$$

  where $P(c)$ is the prior and $\prod_{i=1}^{n} P(x_i \mid c)$ is the likelihood under the assumptions
- **Hyperparameters h:** smoothing factor α, e.g. $\alpha = 1$
- **Parameters θ:** $P(c)$, V (if not given) and $P(w_i \mid c)$ for all $w_i \in V$
- **Parameter estimation (training)**: given the vocabulary $V = \{w_k\}_{k=1}^{K}$ and a training data set $\{(b_1, y_1), \cdots, (b_N, y_N)\}$, where each b_j contains a list of words. Let $N_c = count(y_j = c)$.
  - Likelihood $P(w_i \mid c)$ for each w_i:

$$P(w_i \mid c) = \frac{count(\forall w_i \in b_j \ for \ y_j = c) + \alpha}{count(\forall \ words \in \ class \ c) + \alpha K}$$

  - Prior $P(c)$:

$$P(c) = \frac{N_c}{N}$$

Gaussian naive Bayes classifier (l5)

- **Prediction y:** categorical data $y \in \{1, \cdots, C\}$
- **Variables x_i, $i = 1, \cdots, d$:** continuous numerical data $x_i \in \mathbb{R}$
  - **Assumption:**
    * x_i's are independent - **NAIVE!**
    * x_i follows a Gaussian distribution
- **Model g:**

$$\begin{aligned} \hat{y} &= g(x_1, \cdots, x_d) \\ &= \underset{c \in \{1, \cdots, C\}}{\arg\max} P(c) \prod_{i=1}^{d} f_i(x_i \mid y = c) \end{aligned}$$

  where $P(c)$ is the prior and $\prod_{i=1}^{d} f_i(x_i \mid y = c)$ is the likelihood under the assumptions with $f_i(x_i \mid y = c) = \frac{1}{\sqrt{2\pi\sigma_{c,i}^2}} e^{-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}}$
- **Parameters θ:** $P(c)$, $\mu_{c,i}$, $\sigma_{c,i}$ in $f_i(x_i \mid y = c)$ for all c and i
- **Parameter estimation (training)**: given a training data set $\{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_N, y_N)\}$, where each $\boldsymbol{x}_j = [x_1^j, \cdots, x_d^j]$ is a vector containing all the features for one data point. Let $N_c = count(y_j = c)$.
  - $\mu_{c,i}$, $\sigma_{c,i}$ in the likelihood $f_i(x_i \mid y = c)$ for all variable i and all classes c:

$$\hat{\mu}_{c,i} = \frac{1}{N_c} \sum_{t=1}^{N_c} x_i^t$$

$$\hat{\sigma}_{c,i} = \sqrt{\frac{1}{N_c - 1} \sum_{t=1}^{N_c} (x_i^t - \hat{\mu}_{c,i})^2}$$

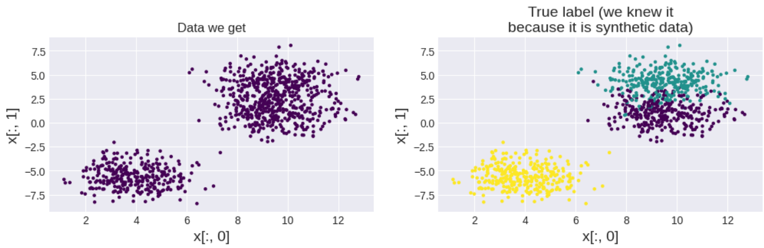
    for all $t \in$ class c
  - Prior $P(c)$:

$$P(c) = \frac{N_c}{N}$$

Machine learning: clustering

K-means (l9)

- **Prediction** y: categorical data $y \in \{1, \cdots, K\}$
- **Variables** x: d dimensional feature vector $\boldsymbol{x}$



- **Model**:

$$y = \arg \min_{k \in \{1, \cdots, K\}} dist(\boldsymbol{x}, \boldsymbol{\mu}_k)$$

  where $dist(\cdot, \cdot)$ is a distance measure; in this course, we use the Euclidean distance; it is **hard clustering** - one data point is assigned to only one cluster
- **Hyperparameters**: K
- **Parameters**: K centroids
- **Parameter estimation**: an iterative method to update the centroids until convergence
  - Randomly choose K centroids $\boldsymbol{\mu}_k$ for $k = 1, \cdots, K$, e.g. randomly choose K data points from $\mathcal{X}$
  - Repeat the two steps below until convergence, e.g. $\boldsymbol{\mu}_k$ does not change anymore
    * For all $i = 1, \cdots, N$, assign $\boldsymbol{x}_i$ to a cluster $\hat{k}_i$ by computing

$$\hat{k}_i = \arg \min_{k \in \{1, \cdots, K\}} dist(\boldsymbol{x}_i, \boldsymbol{\mu}_k)$$

    * Let $\mathcal{X}_k$ be the set of all $\boldsymbol{x}_i$ assigned to cluster k and N_k be the size of $\mathcal{X}_k$, compute

$$\boldsymbol{\mu}_k \leftarrow \frac{1}{N_k} \sum_{\boldsymbol{x}_j \in \mathcal{X}_k} \boldsymbol{x}_j$$

Gaussian Mixture Models (l10, l11)

- **Prediction** y: y can be a set of continuous numerical data K posterior probabilities or categorical data $y \in \{1, \cdots, K\}$
- **Variables** $\boldsymbol{x}$: a d dimensional feature vector $\boldsymbol{x} = [x_1, \cdots, x_d]$ with PDF $f(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{x} \mid k)$
- **Model**: for $k = 1, \cdots, K$

$$\overbrace{P(k \mid \boldsymbol{x})}^{posterior} = \frac{\overbrace{P(k)}^{prior} \; \overbrace{f(\boldsymbol{x} \mid k)}^{\substack{likelihood\ of\ k \\ given\ data}}}{\underbrace{\sum_{c=1}^{K} P(c) f(\boldsymbol{x} \mid c)}_{\substack{likelihood\ of\ the\ mixture \\ distribution\ given\ data}}}$$

  It is **soft clustering** - $\boldsymbol{x}$ is assigned to **all clusters** with a probability - the posterior $P(k \mid \boldsymbol{x})$; **alternatively**, y can be defined as the cluster index with the highest posterior probability, i.e.

$$y = \arg \max_{k \in \{1, \cdots, K\}} P(k \mid \boldsymbol{x}) = \arg \max_{k \in \{1, \cdots, K\}} P(k) f(\boldsymbol{x} \mid k)$$

- **Hyperparameters**: K
- **Parameters**: the parameters of the mixture distribution $f(\boldsymbol{x})$
  - The parameters for each Gaussian likelihood $f(\boldsymbol{x} \mid k)$
  - The prior $P(k)$, typically denoted as π_k
- **Parameter estimation**: the Expectation-Maximization algorithm

Symbols and notations

- Generic mathematical symbol
  - Integral (area under the curve between a and b): $\int_a^b f(x)dx$
  - Summation: $\sum_{i=1}^{N} x_i = x_1 + x_2 + \cdots + x_N$
  - Product: $\prod_{i=1}^{N} x_i = x_1 \times x_2 \times \cdots \times x_N$
  - Factorial: $n! = n \times (n-1) \times \cdots \times 1$
  - Probability of an event: $P(\text{event})$
  - $[a, b]$: the range from a to b, where a and b are numerical values
  - $\{a, b, \cdots\}$: a set that contains elements $a, b, \cdots$
  - Mean value: μ
  - Standard deviation: σ
- Symbols specific in this course
  - N: sample size; number of data points in a data set
  - Chonker duck: a duck that is very round and probably overweight