

Statistical Methods for Data Science: A Starter Kit

Yinan Yu

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

Statistical Data Type

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

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

Data set (a sample): numerical data x_1, \dots, 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:

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

- covariance:

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

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

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

Descriptive Statistics: categorical data

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


- Count/frequency
- Transformed into numerical, discrete data

Visualization: numerical data


- Distribution:


– Histogram/normalized histogram 

– Kernel density estimator 

– Box plot 


- Dependence (two variables):


– Scatter plot 

– Heat map for covariance/correlation 


Visualization: categorical data

- Distribution

– Bar chart 

– Pie chart 

- Dependence

– Mosaic plot 

Probability distribution

- Experiment, sample space, event, probability distribution ($P(\text{event})$ /PDF/PMF/CDF), random variable (discrete/continuous)

- 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

- Bernoulli distribution

- Categorical distribution

- Binomial distribution

- Discrete uniform

- Gaussian distribution

Bayes' rule

- Parameter estimation:

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

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

- Multinomial naive Bayes classifier:

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

- Gaussian naive Bayes classifier/Gaussian mixture models:

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

Q-Q plot

- 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 $p_1, \dots, p_m \in [0, 1]$ (make sure they spread evenly between 0 and 1)
 - For $i = 1, 2, \dots, m$:
 - 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

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

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

Parameter estimation for probabilistic models

- Maximum likelihood estimation: frequentist approach
- Maximum A Posteriori estimation: Bayesian approach

Maximum Likelihood Estimation

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

- Describe the experiments
- Describe the data generated from the experiments
- Describe the random variables (typically with i.i.d. assumption)
- Choose a parameter of interest $\theta \in \mathcal{O}$
- Choose the maximum likelihood estimation as the estimation method:
Given data x_1, \dots, x_N and assume i.i.d. random variables X_i with PDF/PMF $f(x_i)$,

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

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

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

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

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

- 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

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: $f_{X|\Theta}(\mathbf{x} \mid \theta)$ (same as in MLE)
 - Express the posterior distribution in terms of the prior and the likelihood function

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

- 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.