Study guide

Statistics formulae and important details to remember for the exam

Statistic	Formula	Used For
Sample mean (average)	$\bar{x} = \frac{\sum x}{n}$	Measure of center; affected by outliers
Median	n odd: middle value of ordered data n even: average of the two middle values	Measure of center; not affected by outliers
Sample standard deviation	$s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$	Measure of variation; "average" distance from the mean
Correlation coefficient	$r = \frac{1}{n-1}\sum \frac{(x-\bar{x})(y-\bar{y})}{s_x s_y}$	Strength and direction of linear relationship between X and Y

How do you figure out the sample size?

To find the sample size needed to estimate a population mean (μ), use the following formula:

$$n = \left(\frac{z^* \sigma}{MOE}\right)^2$$

In this formula, MOE represents the *desired margin of error* (which you set ahead of time), and σ represents the population standard deviation. If σ is unknown, you can estimate it with the sample standard deviation, s, from a pilot study; z^* is the critical value for the confidence level you need.

What about confidence intervals (CI)?

CI For	Sample Statistic	Margin of Error	Use When
Population mean (μ)	$\bar{x}$	$\pm z^* \frac{\sigma}{\sqrt{n}}$	X is normal, or $n \geq 30$; σ known
Population mean (μ)	$\bar{x}$	$\pm t^*_{n-1} \frac{s}{\sqrt{n}}$	$n < 30$, and/or σ unknown
Population proportion (p)	$\hat{p}$	$\pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$n\hat{p}$, $n(1-\hat{p}) \geq 10$
Difference of two population means ($\mu_1 - \mu_2$)	$\bar{x}_1 - \bar{x}_2$	$\pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	Both normal distributions or $n_1, n_2 \geq 30$; σ_1, σ_2 known
Difference of two population means $\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\pm t^*_{n_1+n_2-2} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$	$n_1, n_2 < 30$; and/or $\sigma_1 = \sigma_2$ unknown
Difference of two proportions ($p_1 - p_2$)	$\hat{p}_1 - \hat{p}_2$	$\pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$n\hat{p}$, $n(1-\hat{p}) \geq 10$ for each group

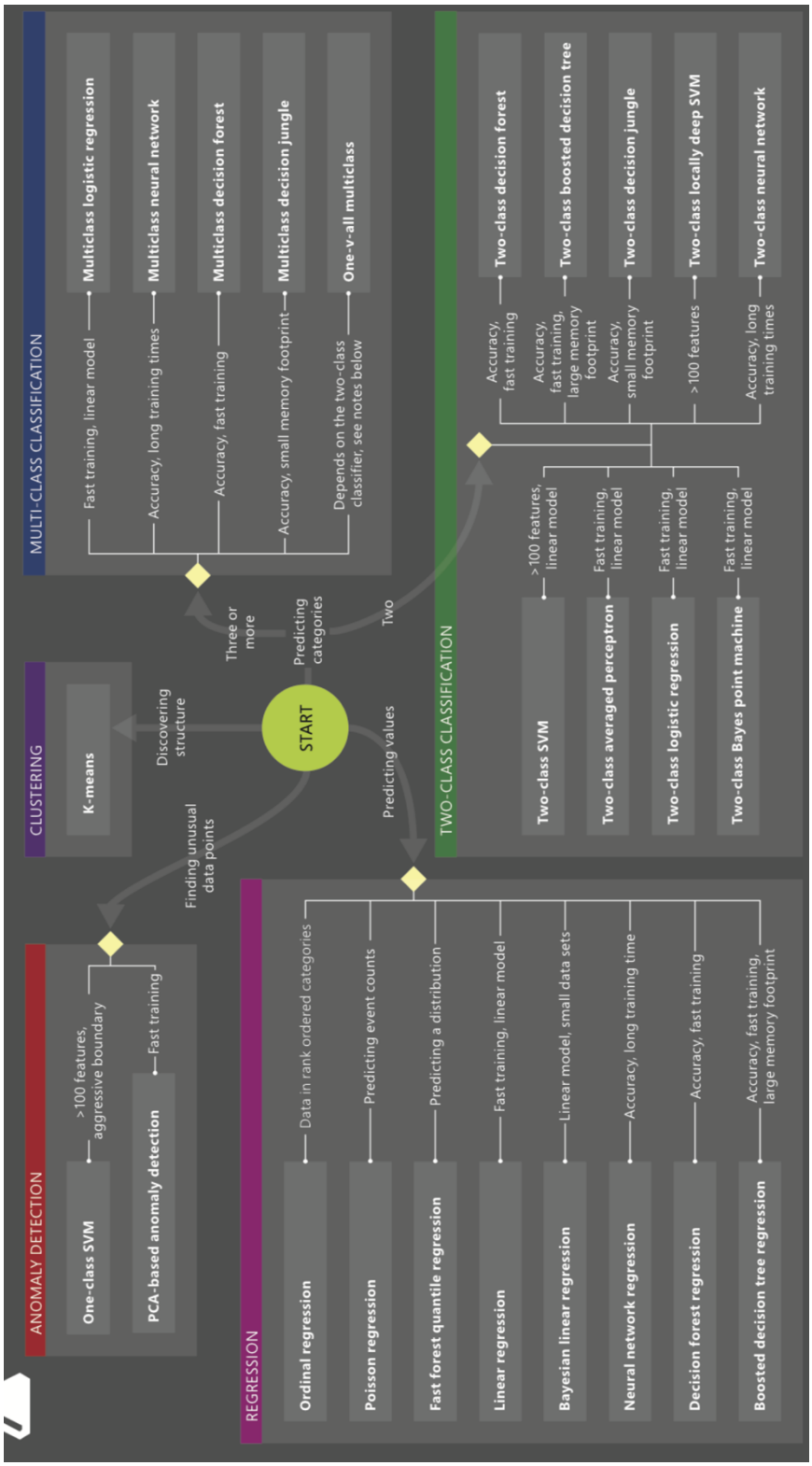
CI critical values: make sure you know the mapping between CI and z*-value (the number of standard errors to be added and subtracted in order to achieve your desired confidence level)

Confidence Level	***z**– value**
90%	1.64
95%	1.96
98%	2.33
99%	2.58

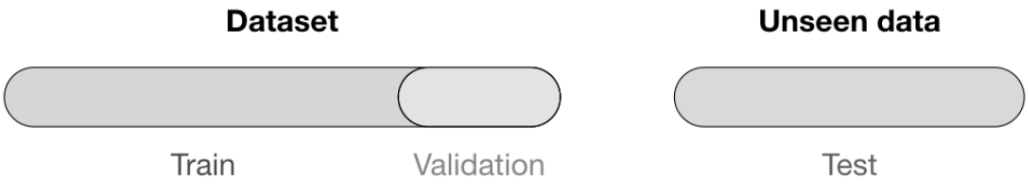
Hypothesis testing

Test For	***Null Hypothesis (H_0)***	***Test Statistic***	***Distribution***	***Use When***
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x}-\mu_o)}{\sigma/\sqrt{n}}$	Z	Normal distribution or $n > 30$; σ known
Population mean (μ)	$\mu = \mu_0$	$\frac{(\bar{x}-\mu_o)}{s/\sqrt{n}}$	t_{n-1}	$n < 30$, and/or σ unknown
Population proportion (p)	$p = p_0$	$\frac{\hat{p}-p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$	Z	$n\hat{p}$, $n(1-\hat{p}) \geq 10$
Difference of two means ($\mu_1 - \mu_2$)	$\mu_1 - \mu_2 = 0$	$\frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	Z	Both normal distributions, or n_1, $n_2 \geq 30$; σ_1, σ_2 known
Difference of two means ($\mu_1 - \mu_2$)	$\mu_1 - \mu_2 = 0$	$\frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	t distribution with df = the smaller of n_1-1 and n_2-1	n_1, $n_2 < 30$; and/or σ_1, σ_2 unknown
Mean difference μ_d (paired data)	$\mu_d = 0$	$\frac{(\bar{d}-\mu_d)}{s_d/\sqrt{n}}$	t_{n-1}	$n < 30$ pairs of data and/or σ_d unknown
Difference of two proportions ($p_1 - p_2$)	$p_1 - p_2 = 0$	$\frac{(\hat{p}_1-\hat{p}_2)-0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$	Z	$n\hat{p}$, $n(1-\hat{p}) \geq 10$ for each group

Remember the different kinds of supervised, unsupervised algorithms, and the differences between classification and regression. This public resource is useful:



Don’t forget about the critical components of model selection: training, validation and testing!



Sample questions on basic machine learning and inference:

What is cross-validation and why must we do it?

What is k-fold cross validation? Are there other kinds of cross validation?

What is the most standard way of evaluating the goodness of fit for a regression? How do you control for adding more and more variables?

What do ROC and AUC stand for? What is their purpose?

What are some ways we can evaluate clustering and other unsupervised methods?