Brandon's Grand Unifying Data Science Map of Everything

LONG FORMAT DATA: It contains values that repeat in the first column. In this format, each row is a one-time point per subject.

WIDE FORMAT DATA: In the Wide Format Data, the data's repeated responses will be in a single row, and each response can be recorded in separate columns.

UNSUPERVISED: THE DATA HAS NO LABELS [NO PRECONCEIVED DESIRED SOLUTION]

SUPERVISED: THE TRAINING DATA INCLUDES THE DESIRED SOLUTIONS [WHICH ARE LABELS ARE 'SPAM' / 'NOT SPAM']

Describe [where do car accidents happen most often?]

Classic Hypothesis Testing

I think the average student is 64 inches tall.

Uo = 64 ← this is the null

Ua > Uo ← this is the alternative

1) The Large-Sample z Test

What is It? A test of a distribution mean μ. The distribution itself is unspecified. (If the distribution is normal, use a t test. See below.)

What is Assumed? The test utilizes the Central Limit Theorem, so the sample size is $n \geq 30$. And all data are sampled from the same distribution, i.e., all data are *independent and identically distributed*, or *iid*.

What is the Test Statistic? The standardized quantity

$$Z = \frac{\bar{X} - \mu_o}{S/\sqrt{n}}.$$

2) The One-Sample t Test

What is It? A test of a *normal* distribution mean μ

What is Assumed? All data are iid samples from a normal distribution of unknown mean and standard deviation.

What is the Test Statistic? The standardized quantity

$$T = \frac{\bar{X} - \mu_o}{S/\sqrt{n}}.$$

3) The Two-Sample t Test

What is It? A test of whether the data in two separate samples were drawn from normal distributions with the same mean μ.

What is Assumed? The data in each sample are normal and are independent and identically distributed. Note that the data may be paired...there may be a one-to-one correspondance between each datum in one sample and each in the other. (This would necessitate a paired t test.)

What is the Test Statistic? Now things start getting a bit complicated. For a classic unpaired two-sample t test, see the formulae on, e.g., this web page.

4) One-Way Analysis of Variance (ANOVA)

What is It? A test of whether the data in k separate samples were drawn from normal distributions with the same mean μ. (It is essentially a k-sample t test.)

What is Assumed? The data in each sample are normal and are independent and identically distributed.

What is the Test Statistic? See this web page.

What is the p-Value? It is output by the `anova()` function in R. See today's lab for more details.

(Wait...is this really useful? It can tell me that one of the means is different...but which one?)

We are glad you asked. You are correct. A standard analysis "protocol" is first to apply ANOVA to see if there is at least one mean that is statistically significantly different from the others (i.e., if $p < \alpha$), then to apply a *post-hoc comparisons test* (like the so-called Tukey test) to see which pairs of means are different. As life is short (as already stated), we will not cover post-hoc tests here.

Classic Summary Statistics

Mean (Average):
- The arithmetic mean, calculated as the sum of all data values divided by the number of data points.
- Formula: $\text{Mean} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Median:
- The middle value in a dataset when it is sorted. It divides the data into two equal halves.
- For an odd number of data points, it's the middle value; for an even number, it's the average of the two middle values.

Mode:
- The value(s) that appear most frequently in the dataset. A dataset can have one mode (unimodal), more than one mode (multimodal), or no mode at all.

Range:
- The difference between the maximum and minimum values in the dataset, indicating the spread of data.
- Formula: $\text{Range} = \text{Max} - \text{Min}$

Variance:
- A measure of the spread or dispersion of data points from the mean.
- Formula for population variance: $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$
- Formula for sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Standard Deviation:
- The square root of the variance, representing the average deviation of data points from the mean.
- Formula: $\text{Standard Deviation} = \sqrt{\text{Variance}}$

Interquartile Range (IQR):
- The range between the first quartile (25th percentile) and the third quartile (75th percentile). It measures the spread of the middle 50% of data points and is resistant to outliers.

Skewness:
- A measure of the asymmetry of the data distribution. Positive skewness indicates a longer tail on the right, while negative skewness indicates a longer tail on the left.

Kurtosis:
- A measure of the "tailedness" of the data distribution. It quantifies whether the data distribution has heavier or lighter tails than a normal distribution.

Correlation Coefficient (Pearson's r):
- A measure of the linear relationship between two variables. It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear correlation.

Distribution Shape

Name	Type	Parameters	Domain (i.e., allowed x values)
Binomial	discrete	trials n, probability of success p	integers 0 to n
Poisson	discrete	expected number of counts λ	all integers ≥ 0
Normal	continuous	mean μ, standard deviation σ	$(-\infty, \infty)$
Standard Normal	continuous	none ($\mu = 0$ and $\sigma = 1$)	$(-\infty, \infty)$
t	continuous	number of degrees of freedom ν	$(-\infty, \infty)$
Exponential	continuous	event rate β	$(0, \infty)$
Chi-Square	continuous	number of degrees of freedom ν	$(0, \infty)$

Visualization

Mapping

Spatial Analysis

Detect [is there a car in this image?]

Generative Modeling for Distributions

Generative Adversarial Networks (GANs)

Variational Autoencoders (VAEs)

Association learning

Anomaly Detection

Structural learning

Clustering

K- Means Cluster

What is the right number of K?

Odds vs Evens

Generally, you should use odd values of k to avoid ties. Ties can occur when the same number of neighbors from each class are present, which can complicate the decision.

K Fold Cross Validation

Evaluate the performance of the model with different values of k.

Grid Search

Perform a search over a range of k values and evaluate the model's performance for each value. Choose the k that gives the best results on the validation set

Elbow Method

Apply K-means over range of k-values from k=1 to k=max. For each k, calculate the Sum of Squares (distance between each data point and assigned cluster - see WSS). Plot (left image) and use your best judgement

$$WSS = \sum_{i=1}^{n}\sum_{j=1}^{k} ||x_i - c_j||^2$$

Optimal number of clusters

Elbow method

Gap Statistic

Pass an argument for the gap statistic and look for the highest value amongst k

`fviz_gap_stat(gs)`

Optimal number of clusters

We're looking for the highest value. That is 2.

- Hierarchical Clustering
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- Agglomerative Clustering
- Mean Shift Clustering
- Affinity Propagation
- BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)
-OPTICS (Ordering Points To Identify the Clustering Structure)
- Kernel K-Means

Predict [How much will this car cost when it sells?]

What are we trying to predict?

Predicting a Category [Is this a truck, a car, or a van?]

Binary Categories [True/False, Yes/No, This/Not this]

Logistic Regression

Support Vector Machine

Decision Tree

Random Forest

K-Nearest Neighbor

Gradient Boosting

Naive Bayes

Remember that statistical models do not predict a class but rather estimate a probability that a datum belongs to class 0 or 1. As a rule of thumb, if your classes are imbalanced (say 60/40) then adjust the threshold t to approximately the proportion of data in class 1, 0.6

Multiple Possible Categories [Risk: High, Medium, Low]

One vs Rest

Decision Tree

Random Forest

K-Nearest Neighbor

Gradient Boosting

ONE V REST - Naive Bayes

ONE V REST - Support Vector Machine

ONE V REST - Logistic Regression

Truly Distinguish

Decision Tree

Random Forest

K-Nearest Neighbor

Gradient Boosting

Have loads and loads of data and need even more accuracy?

Deep Learning

RNN

Predicting a Quantity [How much should this car sell for?]

Decision Tree

Random Forest

Gradient Boosting

Support Vector Machine

K-Nearest Neighbor

Regression? See Inference Section

A good suite of models to explore, in order from least flexible/most inferential to most flexible/least inferential would be

- regression (quantitative response): linear regression, bestglm, regression tree, random forest, extreme gradient boosting, K-nearest neighbors (large amount of data), deep learning (huge amount of data)...pick the model with the lowest test-set mean-squared error
- classification (factor response, two outcomes): logistic regression, bestglm, classification tree, random forest, extreme gradient boosting, K-nearest neighbors (large amount of data), support vector machines (small amount of data), deep learning (huge amounts of data)...pick the model with...well...this is more complicated

Inspect data and check for correlations

Highly Correlated Features? You may have Multicollinearity.

① Check for redundancy in the correlation matrix -> consider removing one of them or combining them

② Try forward selection, backward elimination, or recursive feature elimination to select only the most important features

③ Do a Principal Component Analysis to reduce dimensionality. PCA creates linear combinations of features, potentially reducing multicollinearity.

④ Limit yourself to only regularization methods for regression (Lasso L1 or Ridge L2)

Split Data into Training and Test Set

```
from zlib import crc32

def test_set_check(identifier, test_ratio):
    return crc32(np.int64(identifier)) & 0xffffffff < test_ratio * 2**32

def split_train_test_by_id(data, test_ratio, id_column):
    ids = data[id_column]
```

Imbalanced Classes

Change the decision threshold probability in model

Have lots of data?

Undersampling

SMOTE: Synthetic Minority Over-sampling Technique (creates new examples by interpolating between existing examples of the minority class).

Have moderate amount of data?

Oversampling

When should I sample in general?

Every element of p has equal chance of being afflicted?

Simple Random Sample

Systemic (every 15th person)

Population has natural and meaningful subgroups/strata

Stratified random sample

Too hard to get a complete list of all elements (e.g. US population)

Cluster Sampling (e.g. a city/ a school)

Hard to reach subjects? Hidden population?

Snowballing Sampling (ask a few and then ask for referrals of others who are the same)

Clean data, fill or drop missing values

**Warning: Make sure you split the data BEFORE imputation to avoid bias and data leakage. Set the random seed, save the test set in a variable, or better yet use each UID.

Imputation Options

Missing at Random?

- MCAR: Missing completely at random,
- MAR: Missing at random, or
- MNAR: Missing not at random.

- **MCAR:** Any two individuals, regardless of their values of disease status, level of exposure, and age, have the same probability of having a missing value for exposure.
- **MAR:** Any two individuals with the same disease status and age have the same chance of having a missing exposure value, regardless of how large or small their actual level of exposure.
- **MNAR:** Even among individuals with the same disease status and age, the chance of having a missing exposure value depends on their level of exposure.

Column Mean or Median (or even Mode)

Normally distributed?

Hot Deck Imputation

KNN Imputation

Time series? / Longitudinal data?

Last Observation Carried Forward

Backfill/ Forwardfill

Strong relationship with other variables?

Regression Impute

Have text or categorical data?

Hierarchical Categories (Good, Average, Bad) --> Ordinal Encoder

Regular Categories (Ocean-Front, Mountain, Plains) --> One Hot Encoder

Have hundreds or more categories?

Implement Sparse Matrix (SciPy rather than Numpy Array)

Unstructured Text that is not a category?

Long descriptions (more than a few hundred words)?

Short descriptions (2-3 paragraphs or shorter)?

Term Frequency-Inverse Document Frequency (TF-IDF)

Topic Modelling

Latent Dirichlet Allocation (LDA) or Non-negative Matrix Factorization (NMF)

Named Entity Recognition

Sentiment Analysis

Feature Scaling

With few exceptions, Machine Learning algorithms don't perform well when the input numerical attributes have very different scales. This is the case for the housing data: the total number of rooms ranges from about 6 to 39,320, while the median incomes only range from 0 to 15. Note that scaling the target values is generally not required.

Min-Max

Values are shifted and re-scaled to range from 0 to 1. Subtract the min value and dividing by the max minus the min.

Standardization

Subtract the mean value and then divide that number by the standard deviation

- Con: Does not usually work with Neural Networks
- Pro: Much less affected by outliers.

Suppose a district had a median income equal to 100,000. Min-max scaling would then crush all the other values from 0–15 (most incomes) down to 0–0.15, whereas standardization would not be much affected.

$$X \rightarrow \frac{X - \bar{X}}{S_X},$$

Have too many features? Not sure which is good?

Filter Methods

Relatively easy but not as robust

ANOVA / Chi Square

Linear Discrimination

Wrapper Methods

Resource Intensive but accurate

Forward/ Backward Selection

Recursive Feature Elimination

Best Subset Selector: Try every single combination and show the best one. Since linear regression is so fast, you can usually do an exhaustive search. BSS works best for less than 25-30 predictors.

BSS output is:

$$C_p = \frac{1}{n}(\text{RSS} + 2k\hat{\sigma}^2)$$
$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}(\text{RSS} + 2k\hat{\sigma}^2) = \frac{C_p}{\hat{\sigma}^2}$$
$$\text{BIC} = \frac{1}{n}(\text{RSS} + \log(n)k\hat{\sigma}^2)$$

Worried about underfitting?

AIC: you can be confident that your selected variables include all of the important ones, but the final list might include a handful of unimportant ones as well. Tends to overfit

Worried about overfitting?

BIC: you can be confident that every selected variables are important but other important variables might be left out of the final list. Tends to underfit

Best GLM

Select and Train Models

Evaluate Models based on Metric

Regression	Classification	Recommender System
Mean Absolute Error (MAE) Root Mean Squared Error (RMSE) R-Squared and Adjusted R-Squared	Recall Precision F1-Score Accuracy Area Under the Curve (AUC)	Mean Reciprocal Rank Root Mean Squared Error (RMSE)

K-Fold Cross Validation

** not always possible given data size*

```
from sklearn.model_selection import cross_val_score
scores = cross_val_score(tree_reg, housing_prepared, housing_labels,
                         scoring="neg_mean_squared_error", cv=10)
tree_rmse_scores = np.sqrt(-scores)
```

Youden's J Statistic: attempts to find the threshold where the sum of the correct classifications for both classes are maximized. Sacrifices accuracy in the dominant class to do well in both.

By introducing ROC curves and the concept of AUC, we now have a new metric for model selection. But ultimately, we do need to choose *one* threshold value so as to generate predictions. We can use the ROC curve to do this, while acknowledging that there is no unique means to choose that one value.

An often-used choice is *Youden's J* statistic:

$$J = \text{sensitivity} + \text{specificity} - 1.$$

Basically, we want to determine the threshold value that allows us to best identify NON-CBs and CBs at the same time. For our example:

```
J = roc.log$sensitivities + roc.log$specificities - 1
w = which.max(J)
cat("Optimum threshold for logistic regression: ",round(roc.log$thresholds[w],3),"\n")
```

```
## Optimum threshold for logistic regression:  0.488
```

The value is close to 0.5, which is not surprising because our data had balanced classes (i.e., just as many CBs as NON-CBs). When the data feature unbalanced classes, optimal thresholds may differ greatly from 0.5.

Problem Statement

Binary Classification (A,B = 0.5)

Class Labels

Probability

Metrics

1. ROC Curve
2. AUC Curve
3. PR Curve

Balanced Dataset

Metrics

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Imbalanced Dataset

Metrics

1. Precision
2. Recall
3. F Beta

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

Confusion Matrix

$$\text{Type I Error} / \text{FPR} = \frac{FP}{FP + TN}$$

Type II Error

Always have to focus on reducing the Type I Error and Type II Error

Recall (TPR) : Out of the total positive actual value, how many values did we correctly predicted positively. Whenever the FN is important wrt the problem statement, use Recall.

Precision : Out of the total predicted positive results, how many results were actual positive. Whenever the FP is important wrt the problem statement, then use Precision

In some of the problem statements FP and FN both are equally important in an imbalanced dataset.In such cases we have to consider both **Recall** and **Precision.** In such cases we have to use **F-beta score**

$$\text{F-Beta} = \frac{(1+B^2)\ \text{Precision X Recall}}{B^2 \text{ x (Precision + Recall)}}$$

1. If we have a problem statement where the **FP** and **FN** are both equally important, we select **Beta = 1**
2. In scenarios where the **FP** is having greater impact than the **FN**, in such case we reduce the **Beta** value between 0 to 1
3. In scenarios where the **FN** is having greater impact than the **FP**, in such case we increase the **Beta** value greater than 1 and might range between 1 to 10

Shortlist the best performing models

Fine-Tune them

Grid Search

```
from sklearn.model_selection import GridSearchCV

param_grid = [
    {'n_estimators': [3, 10, 30], 'max_features': [2, 4, 6, 8]},
    {'bootstrap': [False], 'n_estimators': [3, 10], 'max_features': [2, 3, 4]},
  ]

forest_reg = RandomForestRegressor()

grid_search = GridSearchCV(forest_reg, param_grid, cv=5,
                           scoring='neg_mean_squared_error',
                           return_train_score=True)

grid_search.fit(housing_prepared, housing_labels)
```

Rather than print a bunch of stuff just ask for the best combo of arguments....

```
>>> grid_search.best_estimator_
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
           max_features=8, max_leaf_nodes=None, min_impurity_decrease=0.0,
           min_impurity_split=None, min_samples_leaf=1,
           min_samples_split=2, min_weight_fraction_leaf=0.0,
           n_estimators=30, n_jobs=None, oob_score=False, random_state=None,
           verbose=0, warm_start=False)
```

Too many parameters? Taking too long?

Random Search

when the hyperparameter search space is large, you'll have to use RandomizedSearchCV instead.

Ensemble Methods

Make a table of results and select the best model

Continuously test model against new observations

Optimize [how many of each type of car should I manufacture?]

Time Series Forecssting

Survival Analysis

Continuous Optimization

Queing

Stochastic Optimization

Discrete Optimization

Simulation

Linear Programming

Sensitivity Analysis

Integer Programming

Goal Programming

Network Science

How to measure distance?

Eigenvector

Euclidean

Infer [how can I be sure that my policy is having an effect?]

Casual Inference

Instrumental Variables

Attempting to estimate the causal effect of variable X ("explanatory variable") on Y ("dependent variable"), an instrument is a third variable Z affects Y only through its effect on X.

Scientist wants to estimate the effect of smoking (X) on health (Y). Correlation between smoking and population health does not necessarily imply smoking CAUSES bad health (could be because of drinking or depression or being illiterate or extra terrestrial contact). It is impossible to conduct a controlled lab experiment. So the scientist estimates the effect of smoking on health from observation data using the tax rate for tobacco products as an instrumental variable (Z). The tax rate for tobacco products is a reasonable choice for an instrument because the researcher assumes that it can only be correlated with health through its effect on smoking. If the researcher then finds tobacco taxes and state of health to be correlated, this may be viewed as evidence that smoking causes changes in health.

Regression

Ordinary Least Squares

Check Assumptions

Assumptions are never perfectly met, but they should be reasonable

Linear: A change in the independent variable is proportional to the change in the response variable

Homoscedascity: the variance among errors is constant; consistent noise among the data

Not sure if it is linear?

1. Create scatter plots of the independent variable(s) against the dependent variable. Look for a roughly linear relationship. Points should cluster around a straight line

2. Plot the residuals (the differences between the observed and predicted values) against the fitted values (the predicted values). The plot should not show a clear pattern; it should be evenly distributed around the horizontal line at zero.

Spread of the residuals not consistent?

Breusch-Pagan Test

This is a statistical test that checks for heteroscedasticity in the residuals. A significant p-value suggests that heteroscedasticity may be present

Figure 4.1 *Residuals vs. fitted plots—the first suggests no change to the current model while the second shows nonconstant variance and the third indicates some nonlinearity, which should prompt some change in the structural form of the model.*

Heteroscedascity? (stabilize the variance and to bring the distribution closer to the Normal distribution)

Try using non-linear transformations

log-transform for variables with exponential growth).

Taking the square root can address right-skewness

Heteroscedastic Robust Standard Errors: technique to obtain unbiased standard errors of OLS coefficients under heteroscedasticity

Weighted Least Squares (WLS): assigns different weights to observations based on their estimated variances.

Non-linear?

Remove outliers

Did the residual standard error (RSE) drop significantly?

Good; Recheck assumptions and continue with regression

Regression Spline

Local Polynomial regression

Warning: individual regression coefficients for the regression-spline components are uninterpretable

Independent Errors: the residual (errors) for one observation should not provide info about error of another observation

No auto-correlation: The observations are independent of each other (time series data)

Think you may have auto correlation?

Durbin Watson Test

Takes values between 0 and 4. A value around 2 indicates no autocorrelation. Values significantly different from 2 may suggest autocorrelation.

Plot Residuals vs Time

Plot the residuals against the order or time at which they were observed. There should be no discernible pattern or trend in the plot, indicating independence.

Ljung-Box Test

Tests for significance for autocorrelation at lags up to a specified number. A low p-value suggests the presence of autocorrelation.

Do you have access to the same data at specific and regular intervals?

Time Series Analysis

Normal Errors: Errors are normally distributed

Errors not following a normal distribution?

Quantile Quantile Plot or Probability Plot (Q-Q Plot)

Plot the standardized residuals against theoretical quantiles from a normal distribution. If the points closely follow a straight line, it suggests that the residuals are normally distributed.

Shapiro-Wilk Test

Test for normality: The null hypothesis is that the errors are normally distributed. A high p-value suggests that the data does not significantly deviate from normality.

Transform response variable - log, square root, Box-Cox transformation

The Box-Cox transformation is defined as:

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y) & \text{if } \lambda = 0. \end{cases}$$

Consider a Generalized Linear Model (if response variable follows Poisson, Gamma, etc). GLMs allow for modeling relationships with non-Gaussian error distributions.

Check for outliers

No multicollinearity: There should be no high correlation among independent variables.

No perfect colinearity: there should not be a perfect linear relationship among some or all of the predictors.

Create correlation matrix as you normally do

Seeing high numbers?

VIF (Variance Inflation Factor)

Measures how much the variance of an estimated regression coefficient increases when your predictors are correlated. A high VIF (typically > 10) suggests multicollinearity.

Auxiliary regression

Run one X on all the other X's and test hypothesis with formula
Test statistic: F=[R2/(k-1)] / [(1-R2)/(n-k)]

Present?

Lasso Regression
Ridge Regression
Elastic Net Regression

Principal Component Analysis can mitigate multicollinearity if removal is not an option

No omitted variables

Concerned about feasibility?

Do you have access to Panel Data?

Consider longitudinal study

Evaluating Regresion

Elasticity		Marginal Effect
* 1. Linear: $B_2 * X/Y$	$Y_i = B_1 + B_2 X_i + u_i$	B_2
* 2. Quadratic: $(B_2+2\beta_3 X)*X/Y$	$Y_i = B_1 + B_2 X_i + + B_3 X_i^2 + u_i$	B_2+2b_3X
* 3. Double-Log: B_2	$\ln Y_i = B_1 + B_2 \ln X_i + u_i$	$B_2 * Y/X$
4. Linear-Log: $B_2 * 1/Y$	$Y_i = B_1 + B_2 \ln X_i + u_i$	$B_2 * 1/X$
* 5. Log-Linear: $B_2 * X$	$\ln Y_i = B_1 + B_2 X_i + u_i$	$B_2 * Y$
6. Reciprocal: $-B_2 * 1/XY$	$Y_i = B_1 + B_2\ 1/X_i + u_i$	$-B_2 * 1/X^2$
7. Log Reciprocal: $B_2 * 1/X$	$\ln Y_i = B_1 + B_2\ 1/X_i + u_i$	$B_2 * Y/X^2$
* 8. Interaction: $(B_2+B_4Z)*X/Y$	$Y_i = B_1 + B_2 X_{2i} + + \beta_3 Z_i + \beta_4 X_i Z_i + u_i$	B_2+B_4Z