**Probability Theory**

Probability distribution is an equation that links each outcome of a statistical experiment with its probability of occurrence.

f(x | theta) Marginal probability distribution is where f is the probability density of x for all possible values of y, given the distribution parameters. It’s determined from the joint distribution of x and y by integrating over all values of y.

f(x, y | theta) Joint probability distribution is the probability of two or more things happening together. f is the probability of x and y together as a pair, given the distribution parameters. Not independent.

f(x | y, theta) Conditional probability distribution is where f is the probability of x by itself, given specific value of variable y and the distribution parameters theta. P(A|B)=P(AB)/P(B).

Independence is for two events A and B, occurrence of one event doesn’t affect occurrence of another event in no way. P(A and B)=P(A)P(B). P(A|B)=P(A).

Marginal independence is knowing event B doesn’t help in value of event A. P(A=a|B=b1) = P(A=a|B=b2) = P(A=a).

Conditional independence is P(A|B, C)=P(A|C). A is conditionally independent of B given C.

Multiplication rule is P(A, B)=P(A)P(B|A)=P(A|B)P(B).

Derive Bayes Rule is P(B|A)=(P(A|B)P(B))/P(A)

Random variables are function that map from a space of outcomes, known as a *sample space* Ω, to values of a variable *X*. Outcomes are assumed to be the result of a random phenomenon.

How can random variables be used in typical ds application??

Random variables can be used as functions on outcomes in typical ds application.

Probabilistic programs are programs written in Turing-complete programming languages that have syntax and semantics.

Probabilistic programs represent probability distributions.

Program executions represent samples from that distribution.

**Data Science Goals**

Central dogma of data science is: an objective world samples data, use data to generalize models, use models to infer predictions, use predictions to perform actions.

What can go wrong??

Descriptive analytics focus on describing and characterizing the data.

Predictive analytics focus on inference about and modeling of the distribution from which the data were drawn.

Prescriptive analytics focus on inference about and modeling of the underlying system that generates the distribution from which the data were drawn, often for the purpose of causal inference.

Descriptive analytics aims to provide initial insights into the likely form of the DGP. This often suggests a model family.

Predictive analytics aims for models that reproduce the output of the DGP.

Prescriptive analytics aims for models that reproduce key internal structure and parameters of the DGP.

Methods typically used??

Date generating process is a notional process that includes all of the influences that combine to lead to produce individual elements of a sample space.

Statistical model is a formal description that embodies a set of statistical assumptions and represents the DGP.

Statistical models can be represented as programs.

As with DGPs, they are often presented in an abbreviated form and are relatively short, imperative and procedural, and probabilistic.

Common statistical models can all be represented as programs.

But statistical models cannot be represented by common model forms.

Parametric model is a model that can be described using a fixed number of parameters.

Non-parametric model is a model that grow in the number of parameters to accommodate the complexity of the data. Parameters could be infinite.

Neural network, kernel density estimation, k-nearest neighbors are non-parametric.

Simple linear regression, simple Bayes classifier are parametric.

**Data Representation and Sampling**

Data instances group potentially related measurements of an outcome together.

Unit of analysis is the smallest target group of your analysis.

Hierarchical structure??

Non-IID with regular relationship structure among instances:

1d, temporal data (stock price), sequences (letters in text or base pairs in DNA)

2d, image data (computer vision), spatial data (GIS or mapping data)

3d, volumetric (atmospheric temperature)

Non-IID with irregular relations: computer networks, social networks, citation graphs.

Nominal (categorial, discrete, binary), values can be distinguished, but not ranked. (eye color, political party)

Ordinal, ranks are meaningful, but not arithmetic transformations. (letter grade)

Interval, distances between values are meaningful, but not ratios. No altitude is not meaningful. (altitude, temperature in Fahrenheit)

Ratio scales, ratios are meaningful, because the zero point is known and meaningful. (height, weight, degrees Kelvin)

Outliers are observations that lie abnormal distances from other values in a random sample from a population.

Interquartile range, Standard deviation.

Mean is the arithmetic mean, average value of a distribution.

Median is the value that splits distribution in half.

Median is more robust in the presence of outliers compare with mean, less affected by outliers.

Standard deviation is a measure of how spread out numbers are.

Variance is the average of squared differences from the Mean.

Interquartile range is 75th percentile – 25th percentile.

Why use one versus the other??

Tukey’s ladder of power can change the shape of a skewed distribution so that it becomes normal or nearly-normal. It can also help to reduce error variability.

Tukey’s ladder of power is a family of transformations that can be applied to variables measured on interval or ratio scales.

When we attempt to make a non-linear variable linear so that it can be accurately represented by linear regression.

**Descriptive Analytics**

Exploratory data analysis (EDA) is a collection of methods for analyzing data, that focus on visual presentation and interactivity, and that do not require formulation of explicit hypotheses or statistical models.

EDA is not modeling. It focuses on suggesting possible models, not constructing or evaluating them. EDA is logically prior to modeling.

EDA is not presentation graphics. It focuses on rapid creation and interactive manipulation of informative graphics for the researcher, not preparation of final graphics for papers. And many software packages for presentation graphics are poorly suited to EDA.

Maximizing the data-ink ratio in a graphic means the most of ink is used to present actual data.

Small multiples use the same basic graphic or chart to display difference slices of a data set.

It can show rich, multi-dimensional data without trying to cram all that information into a single, overly-complex chart.

Use bar chart or scatterplot for single categorical variable distribution. Use box plot for one interval or ratio scale variable distribution.

Conditional distribution??

Joint distribution??

Box plot parts and meanings.

Kernel density estimator is constructed by summing a set of smaller distributions each of which corresponds to a data point and each of which has total area 1/N.

Relative advantages??

**Models**

Parameter is numeric or symbolic values that determine how a model maps from input to output.

Hyperparameter is a parameter of a model family that controls the capacity of the learned model. Hyperparameters typically control how an entire model is learned.

Example of hyperparameter is the value k of k-nearest neighbor model or bandwidth parameter of a kernel density estimator.

Capacity is the ability of a model to represent complex mappings from inputs to outputs. For example, models that can represent conditional distributions are said to have higher capacity than models that can represent only simpler conditional distributions.  
 Always using the highest capacity model wastes resources.

Two primary components of error for a learned statistical model are bias and variance.

When the model is underfit, bias will be large. When the model is overfit, variance is large. When the model is appropriate fit, bias and variance is traded off.

**Linear Models**

Slope and intercept are necessary and sufficient components, as well as the variance of the error distribution.

Parameters are estimated by minimizing RSS (residual sum of square).

Residual sum of squares (RSS) is sum of the squared differences between the actual Y and the predicted Y.

p-value measures how compatible your data are given the null hypothesis. High value indicates your data are likely with a true null. Low value indicates your data are unlikely with a true null. True null give beta1=0 no relationship between X and Y, otherwise beta1 not equal 0.

Small p-value indicates there is an association between the predictor and response.

Correlation coefficient tells how strong a relationship is between two sets of data. Range from -1 to 1, it tells how linear the dataset is.

Residual plot is a graph shows the residuals on the vertical axis and the independent variable on the horizontal axis. It can tell whether a linear regression model is appropriate for the data or not.

Heteroscedasticity means the variance of error is not equal across the range of the second variable.

Homoscedasticity means the variance of error is equal across the range of the second variable.

Outlier is a point for which yi is far from the value predicted by the model.

Plotting residual graph or studentized residual graph can identify the outliers. (studentized is dividing each residual by the standard error)

If the outlier is same as the mean of model, then it won’t affect the slope but the intercept. If the outlier is far away from the rest and not stay on mean, then it will affect both parameters.

**Simple Bayesian Classifier**

Probability of prior, probability of likelihood, probability of evidence.

SBC = prior\*likelihood / evidence

For real valued x, normal density.

For binary valued x, Bernoulli distribution.

For general nominal valued x, categorical distribution.

Class probability and parameters of the marginal class conditional distribution are learned by maximum likelihood on a sample of training data.

**K-nearest Neighbor Classifier**

K-nearest neighbor classifier stores the training data and classifies new data based on a majority vote of the nearest k training instances to a test instance.

Minkowski distance is a metric in a normed vector space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance.

Shape??

Low bias, high variance, curse of dimensionality, space and time complexity, lack of selectivity.

Bias is consistent error, variance is random error

Neural network, kernel density, k nearest neighbors, non-parametric, low bias, high variance

Linear model, simple Bayes classifier, parametric, high bias, low variance

Whether the equation can exactly represent the model

Cor tells how linear the dataset is