Module 1 Intro To ML

— Parametric vs non-parametric methods

- Parametric tests are test that make assumptions about the parameters of the population distribution from which the sample is drawn.
- Non-parametric tests are "distribution-free" and, as such, can be used for non-normal variables.

— Regression vs classification

Classification is about predicting a label and regression is about predicting a quantity

Classification = label or class

Regression = quantity

— Nominal, Ordinal, Interval, Ratio

student ID, zipcode — Nominal, Ordinal

team Rank — Ordinal

temperature — Interval

weight, height, Distance — Ratio

Qualitative / Categorical — Quantitave

Ratio	Has absolute ze
Interval	Distance makes sense
Ordinal	Attributes can be ranked
Numinal	Values are names or labels

— Assessing Accuracy

- Confusion Matrix

Actual \ Predicted	+	−
+	True Positive TP	False − FN
−	False + FP	True − TN

- Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$ 90% cancer Free example. not good enough.
- Precision % of + / total Predicted (+'s) $\frac{TP}{TP+FP}$

Opposites of each other

- Recall / sensitivity / True positive Rate $\frac{TP}{TP+FN}$
- specificity % of − / total predicted (−'s) $\frac{TN}{TN+FP}$

- Bias-variance tradeoff - ~~Email spam filtering example~~
  - is the property of a model that the variance of the parameter estimates across samples can be reduced by increasing the bias in the estimated parameters.
  - the amount the expected model prediction differs from the true value of target or how far off our predictions are from real values

- Flexable vs inflexable
  An inflexible model will perform better in general, a flexable model will cause overfitting because of the small sample size. This means a bigger inflation in variance and a small reduction in bias. A flexable model will capture too much of the noise in the data due to the large variance of the errors.

- Bayes Optimal Classifier
  Spam email example.
  split into 2 groups Normal + Spam
  Count words Probability of email being normal or Spam
  "Dear" "Friend" "lunch" "Money" +1 to all words so you don't get "0"

- KNN
  non-parametric - classification and regression
  if K=1 then the object is simply assigned to the class of that single nearest neighbor.

Module 2: Getting Started with R.

Basics of R and RStudio

Vector, matrix, array, dataframe, and list

Control structures: sequence, selection, and iteration.
Know R function to manipulate summarize and explore datasets.

Module 3: Linear Regression

- Explanatory vs predictive modeling
  predictive modelling is "what is likely to happen?"
  Explanatory modelling is all about "what can we do about it?"

- Regression coefficient estimation
  Regression coefficients are estimates of the unknown population parameters and describe the relationship between a predictor variable and the response. In Linear Regression, coefficients are the values that multiply the predictor values.

- Null and Alternative hypothesis

  A statement about the value of a population parameter in case of 2 hypotheses, the statement assumed to be true is called the null hypothesis (H_0) and the contradictory statement is called the alternative hypothesis (H_a)

- P-value

  if > 0.05 item is not statistically significant

  if $<= 0.05$ item is statistically significant.

- Qualitative Predictors

  - Differences in credit card balance between male and females.

Qualitative Predictors with 2 levels

$$\text{female}_i = \begin{cases} 1 & \text{if the ith Person is female} \\ 0 & \text{if the ith Person is male} \end{cases}$$

Qualitative Predictors with more than 2 levels

$$\text{ethnicity\_asian}_i \begin{cases} 1 & \text{if the ith person is asian} \\ 0 & \text{is not asian} \end{cases}$$

$$\text{ethnicity\_Caucasian}_i \begin{cases} 1 & \text{is Caucasian} \\ 0 & \text{is not Caucasian} \end{cases}$$

- Interaction terms in linear regression

in marketing synergy effect is spending 50k on TV and 50k on radio advertising resulting in more sales than allocating 100k to only one.

$$Sales = B_0 + B_1 \times TV + B_2 \times Radio + B_3 \times TV \times Radio + \varepsilon$$

- non-linear fit w/ polynomial regression.

$$y = B_0 + B_1 X + B_2 X^2 + \dots + B_n X^h + \varepsilon$$

h = 2 quadratic
h = 3 cubic
h = 4 quartic

Polynomial regression is a linear model, since the outcome y is a linear combination of coefficients B_i $(i = 1, 2, \dots h)$

The hierarchy Principle:

If the polynomial regression model contains X^h and its coefficients is significant, the the model should also include all lower-degree terms X^j $(j<h)$, no matter those X^j are significant or not.

- Polynomial Regression may be misleading if you don't have a large dataset.
- Do not extrapolate beyond your ~~observation~~ observed data range.

— Common Problems using linear regression

(1) Non-linearity of the response-predictor relationships.
(2) Correlation of error terms
(3) Non-Constant variance of error terms (heteroscedasticity)
(4) Outliers
(5) High-leverage points
(6) Collinearity

— Linear Regression vs KNN

K is large, KNN is a little worse than linear Regression
K is small, KNN performs far worse.
KNN suffers from the curse of dimensionality.

Module 4: Classification

$y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{drug overdose} \\ 3 & \text{seizure} \end{cases}$

ordinal Nominal

$y = \begin{matrix} 1 = yes \\ 0 = no \end{matrix}$ => linear Probability Model

if $\hat{y} > 0.5$ = yes

values outside [0, 1] interval become hard to predict.

$$\text{Accuracy} = \frac{NN + YY}{\text{total}}$$

$$\text{sensitivity} = \frac{YY}{T_1 Y}$$

$$\text{specificity} = \frac{NN}{T_1 N}$$

	N	Y	T_2
N	NN	NY	$T_2 N$
Y	YN	YY	$T_2 Y$
T_1	$T_1 N$	$T_1 Y$	$T_1 T_2$ = Total

$$\text{False} + = \frac{YN}{T_1 N} = 1 - \text{specificity}$$

$$\text{False} - = \frac{NY}{T_1 Y} = 1 - \text{sensitivity}$$

$$\text{Balanced Accuracy} = \frac{\text{sensitivity} + \text{specificity}}{2}$$

Resample the training data Deal with imbalanced Data
over-sample the minority class
under-sample the majority class

- use different threshold for prediction
- Customize your cost function to assign larger penalty to the misclassified minority class

Use 0.2 as threshold for prediction

	0.5		0.2
False + =	0.2%	↗	2.4%
False - =	76.3%	↘	42.0%

Good classifier has large area under curve (AUC)

.90-1 = A
.8-.9 = B
.7-.8 = C
.6-.7 = D
.5-.6 = F

LDA tends to be better if there are relatively few training observations so reducing variance is crucial.

QDA is recommended if the training set is very large, so that the variance of the classifier is not a major concern. or if the assumption of a common covariance matrix for the K classes is clearly untenable.