Data Science Specialization in Coursera

Cheng

Friday, February 06, 2015

Contents

Introduction

The purpose of this page is to collect the tools, knowledge, information, methods and concepts that relate to the data science. The basic structure is following to data science specialization in coursera.

In additiion, I roughly divide it into three entries namely Design, methods and product development. Research Design focuses on the questions like how to ask a right question, how to get a solid conclusion, what's genergal approach to evaluate a study and so on. The Reasearch Method portion itemizes the tools on visualizations, statistics and machine learnings. The format of itemization would be giving both the basic procedures and R examples (probally python examples in the future). The last portion Product development contains the information that is not direct related to data science research but more to the rendering process of the research product like reprots and webpages.

Research Design

Formulating a problem

- Question, Modeling, and Validation (QMV) process of data analysis
  - turn a vague question into a statistical one that can be analyzed with statistics and machine learning
  - build rigorous mathematical, statistical, and machine learning models so you can make accurate predictions
  - use fundamental and important metrics that you can use to grade the performance of the models that you've build

Research Methods

Analytic Graphics   Principles

- Comparisons
- Causality, mechanism, explanation, systematic structure
- Multivariate data
- Integration of evidence
- Describe and document the evidence with appropriate labels, scales, sources
- Quality, relevance and integrity of the content

Purpose

- To understand data properties
- To find patterns in data
- To suggest modeling strategies
- To “debug” analyses
- To communicate results

Simple Summary Technics

- One Dimension:
  - five-number summary
  - Boxplots
  - histograms
  - density plot
  - Barplot
- Two Dimensions:
  - Multiple/Overlayed 1-D plots
  - scatterplots
  - smooth scatterplots
- More than Two Dimensions:
  - Multiple/Overlayed 2-D plots or coplot
  - use color, size, shape to add dimensions
  - spinning plots
  - 3-d plots
  - pair plot

Plotting Tools in R

- base plotting system
- ggplot2 system
- lattice system

Clustering
Clustering organizes things that are close into groups.

- How do we define close -> Distance or simliarity.
  - Pick the one that make sence to your problem.
  - Continuous: euclidean distance, correlation similarity
  - Binary: manhattan distance
- How do we group things?
  - Hierarchical Clustering
    | `A agglomerative approach`: 1) Find closest two things; 2) Put them together; 3) Find next
    | `Requires`: 1) A defined distance; 2) A merging approach.
    | `Procedures`: A tree showing how close things are to each other.
    | `R example`: here
    | `Note 1`: picture may be unstable | `Note 2`: determistic
  - K-means Clustering
    | `A partioning approach`: 1) Fix a number of clusters; 2) Get "centroids" of each cluster; 3) Assign things to the closest centroid; 4) Recalculate centroids
    | `Requires`: 1) A defined distance metric; 2) A number of clusters. 3) An initial guess as to cluster centroids
    | `Procedures`: Final estimate of cluster centroids and an assignment of each point to clusters
    | `R example`: here
    | `Note 1`: Require a number of clusters by eyes or cross validation. more
    | `Note 2`: Not deterministic caused by different # of clusters and iterations
  - Dimension Reduction- SVD
    | $X = UDV^T$
    | U(left sigular vector) and V(right sigular vector) are orthogonal,D(sigular value) is a diagonal matirx
    | `R example`:here | `Alexander Ihler video on Youtube`:here
  - Dimension Reduction- PCA
    | $X^TX = VDV^T = UDU^T$
    | `Alexander Ihler video on Youtube`:here
    | `Jason Liu on Quora`:here
- How do we visualize the grouping
  - Dendrogram
    | `R example`: 1,2
  - Heatmap
    | `R example`: 1,2
  - Multidimenional Scaling
    | `R example`: 1

Statistical Inference Here I only put part of concepts and fomulas that are important or easy to remember. When looking the deviation or uitility of R codes, please refer back to the slides linked to each subjuct.

- Paramount among our concerns are:
  - Is the sample representative of the population that we'd like to draw inferences about?
  - Are there known and observed, known and unobserved or unknown and unobserved variables that contaminate our conclusions?

  - Is there systematic bias created by missing data or the design or conduct of the study?
  - What randomness exists in the data and how do we use or adjust for it? Here randomness can either be explicit via randomization or random sampling, or implicit as the aggregation of many complex uknown processes.
  - Are we trying to estimate an underlying mechanistic model of phenomena under study?
  - Statistical inference requires navigating the set of assumptions and tools and subsequently thinking about how to draw conclusions from data.

- Example goals of inference
  - Estimate and quantify the uncertainty of an estimate of a population quantity (the proportion of people who will vote for a candidate).
  - Determine whether a population quantity is a benchmark value ("is the treatment effective?").
  - Infer a mechanistic relationship when quantities are measured with noise ("What is the slope for Hooke's law?")
  - Determine the impact of a policy? ("If we reduce polution levels, will asthma rates decline?")

- Example tools of the trade
  - Randomization: concerned with balancing unobserved variables that may confound inferences of interest
  - Random sampling: concerned with obtaining data that is representative of the population of interest
  - Sampling models: concerned with creating a model for the sampling process, the most common is so called "iid".
  - Hypothesis testing: concerned with decision making in the presence of uncertainty
  - Confidence intervals: concerned with quantifying uncertainty in estimation
  - Probability models: a formal connection between the data and a population of interest. Often probability models are assumed or are approximated.
  - Study design: the process of designing an experiment to minimize biases and variability.
  - Nonparametric bootstrapping: the process of using the data to, with minimal probability model assumptions, create inferences.
  - Permutation, randomization and exchangeability testing: the process of using data permutations to perform inferences.

- Probability
  - keywords:probability, random Variables, PMF, PDF, CDF, survival function, quantiles

- Expectations
  - key words:
    | calculation of expected values of discrete random variables or contious random varibales
    | expected value is a linear operator
    | calculate variance and standard deviation
    | Chebyshev's inequality: $P(|x-\mu| >= k\sigma) <= 1/k^2$

- Independence
  - key words:
    | IID
    | join probability of IIDS
    | covariance : $Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)] = E[XY] - E[X]E[Y]$
    | correlation: $Cor(X,Y) = Cov(X,Y)/\sqrt{Var(X)Var(y)}$
    | sample variance S: $S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$
    | variance of sample mean: $Var(\bar{X}) = \sigma^2/n$. σ is the poluation variance.
    | standard error of sample mean: $\sigma/\sqrt{n}$

| estimate population variance by sample variance: S^2 estimates σ^2; $S/\sqrt{n}$ estimates $\sigma/\sqrt{n}$ the standard error of the mean

- Conditional Probability
  - key words:
    | def: $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$
    | bayes' rules: $P(B|A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B)+P(A \mid B^C)P(B^C)}$
    | sensitivity: $P(+ \mid D)$
    | specificity: $P(- \mid D^C)$
    | positive predictive value: $P(D \mid +)$
    | negative predictive value: $P(D^C \mid -)$
    | prevalence of a disease: $P(D)$
    | diagnositic likehood of a positive test(DLR_+): sesitivity/(1- specificity)
    | diagnositic likehood of a negtative test(DLR_+): (1-sesitivity)/specificity
    | likehood ratios: $\frac{P(D \mid +)}{P(D^C \mid +)} = \frac{P(+ \mid D)}{P(+ \mid D^C)} * \frac{P(D)}{P(D^C)}$. post-test odds of D $= DLR_+ *$ pre-test odds of D
- Common Distributions
  - Bernoulli
    | def: Bernoulli random variables take (only) the values 1 and 0 with probabilities of p and (1-p)
    | PMF: $P(X = x) = p^x(1-p)^{1-x}$ respectively
  - multiple iid bernoulli:
    | PMF: $\prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$
    | maximum likelihood estimator for p: $\hat{p} = \sum_i x_i/n$
  - bionomial:
    | def: the sum of iid Bernoulli trials
    | PMF: $P(X = x) = \left(\ n \ \ x\ \right) p^x(1-p)^{n-x}$
  - normal:
    | PDF: $(2\pi\sigma^2)^{-1/2}e^{-(x-\mu)^2/2\sigma^2}$ If X a RV with this density then $E[X] = \mu$ and $Var(X) = \sigma^2$
    | The MLE for μ is $\bar{X}$.
    | The MLE for σ^2 is $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$ (Which is the biased version of the sample variance.)
  - standard normal:
    | def: When $\mu = 0$ and $\sigma = 1$
    | RVs are often labeled Z and The standard normal density function is labeled ϕ | percentiles of standard normal.
  - poisson:
    | PMF: $P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$,$\lambda$ is the mean number of events per unit time
    | accurate approximation to the binomial distribution When n is large and p is smal,$\lambda = np$
- Asymptopia
  - keywords:
    | LLN(Law of Large Numbers): X_i are iid from a population with mean μ and variance σ^2 then $\bar{X}_n$ converges to μ
    | CLT(Central Limit Theorem): the distribution of `averages` of iid variables, properly normalized ($\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\text{Estimate}-\text{Mean of estimate}}{\text{Std. Err. of estimate}}$), becomes that of a standard normal as the sample size increases.
    | confidence intervals of normal distribution: $\bar{X}_n \pm z_{1-\alpha/2}\sigma/\sqrt{n}$
- t Confidence Interval

  - Chi-squared distribution
    | def: $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, S^2 is the sample variance from iid $N(\mu, \sigma^2)$ and $n-1$ degree of freedom
  - Gosset's t distribution
    | def: $\frac{Z}{\sqrt{\frac{\chi^2}{df}}}$
    | confidence intervals of t distribution: $\bar{X} \pm t_{n-1,1-\alpha/2} S/\sqrt{n}$ | note: (1) assume iid normal but it's robust to this assumption (2) roughly symmetric and mound shaped (3) paired observations are often analyzed using the t-interval by taking difference (4) t quantiles become standard normal as large df (5) skewed distribution voiolated t interval assumpltions. (5) In skewed cases, consider taking logs or using different summary like median. (6) for highly discrete data, like binary, other intervals are available.

- Likeklihood

  - def: Given a statistical probability mass function or density, say $f(x, \theta)$, where θ is an unknown parameter, the likelihood is f viewed as a function of θ for a fixed, observed value of x.

  - Interpretation of likelihoods
    | Ratios of likelihood values measure the relative evidence of one value of the unknown parameter to another
    | Given a statistical model and observed data, all of the relevant information contained in the data regarding the unknown parameter is contained in the likelihood.
    | if X_i are independent random variables, then their likelihoods multiply.

- Bayes

  - def: Posterior $\propto$ Likelihood $\times$ Prior
  - beta distribution
    | def: The beta density depends on two parameters α and β $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \quad \text{for} \quad 0 \le p \le 1$
    | The beta distribution is a suitable model for the random behavior of percentages and proportions or between 0 to 1

  - conjugate prior
    | def: if the posterior distributions $p(\theta \mid x)$ are in the same family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function
  - expample of bernoulli likelihood & beta prior

  - highest posterior density(HPD)
    | def: A 95% credible interval, $[a, b]$ would satisfy $P(p \in [a, b] \mid x) = .95$
  - website: bayesian inference

- compare Two independent Groups of normal data

  - pooled variance (S_P)
    | def: $S_p^2 = (n_x - 1)S_x^2 + (n_y - 1)S_y^2/(n_x + n_y - 2)$
    | unbiased estimator of σ^2 in equal variances case
  - equal variances case
    | differnece follows normal distribution: $\bar{Y} - \bar{X} \sim N\left(\mu_y - \mu_x, \sigma^2(\frac{1}{n_x} + \frac{1}{n_y})\right)$
    | or $\bar{Y} - \bar{X} \sim N\left(\mu_y - \mu_x, S_P^2(\frac{1}{n_x} + \frac{1}{n_y})\right)$
    | freedom of t-statistic: $n_x + n_y - 2$
  - Unequal variances case
    | differnece follows normal distribution: $\bar{Y} - \bar{X} \sim N\left(\mu_y - \mu_x, \frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)$ | freedom of t-statistic:

$$\frac{\left(S_x^2/n_x+S_y^2/n_y\right)^2}{\left(\frac{S_x^2}{n_x}\right)^2/(n_x-1)+\left(\frac{S_y^2}{n_y}\right)^2/(n_y-1)}$$

- Hypothesis Testing
  - null hypothesis(H_0):
    | The null hypothesis is assumed true and statistical evidence is required to reject it in favor of a research or alternative hypothesis
  - Type I error - α
    | false positive | incorrect reject of a true H_0
  - Type II error - β
    | false negtive | fail to reject a false H_0
- p Values
  - def:the probability of finding the observed sample results, or "more extreme" results, when the null hypothesis is actually true.
- Power
  - def: the probability of rejecting the null hypothesis when it is false
  - $1-\beta$
- controlling the false positive rate
  - Problem: too many false positives when performing large number of tests
  - Bonferroni correction
    | set your α levels
    | use α_n= α/m
    | call all P-values less than α_n
    | Pros: Easy to calculate, conservative
    | Cons: too Conservative
  - False Discovery rate (FDR) | Suppose you do m tests
    | You want to control FDR at level α so $E\left[\frac{V}{R}\right]$
    | Calculate P-values normally
    | Order the P-values from smallest to largest $P_{(1)},...,P_{(m)}$
    | Call any $P_{(i)} \leq \alpha \times \frac{i}{m}$ significant
- Resampled Inference
  - jackknife
    | def: The jackknife deletes each observation and calculates an estimate based on the remaining of them
    | bias: $(n-1)\left(\bar{\theta}-\hat{\theta}\right)$
    | standard error: $\left[\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_i-\bar{\theta})^2\right]^{1/2}$
  - one bootstrap example
    | Sample n observations with **replacement** from the observed data resulting in one simulated complete data set
    | Take the median of the simulated data set
    | Repeat above two steps times, resulting in simulated medians
    | then calculate mean, standard error and confidence interval of these medians
    | Check more on: An Introduction to the Bootstrap

Regression Models

- Simple Linear Regression
  - Model
    | def: $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i = \sum_{k=1}^{p} X_{ik}\beta_j + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$
    | Fitted response: $\hat{Y}_i = \sum_{k=1}^{p} X_{ik}\hat{\beta}_k$
    | Residuals: $e_i = Y_i - \hat{Y}_i$
    | Residual variance estimate: $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2$
    | Variation of Model: Total Variation = Residual Variation + Regression Variation or $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
    | R^2: the percentage of variation explained by the regression model.$\frac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} = Cor(Y, X)^2$. note: inflate by number of regressors
    | Least squares: minimizes $\sum_{i=1}^{n}\left(Y_i - \sum_{k=1}^{p} X_{ki}\beta_i\right)^2$
    | Coeffients, Intercepts and residuals using R

  - Interpretation of the coeficient | the expected change in the response per unit change in the regressor, holding all of the other regressors fixed

  - Common problems in multivariate simulation
  - Diagnostics
    | Outliers : influence(hatvalues), leverage (cooks.distance)
    | Residuals: normality(Residual QQ plots), Heteroskedasticity(Residual Plots)
    | feature selection: underfiting, more bias; Overfitting, more variation. F-test on nested model explanation, F-test on nested model R example

- Generalized Linear Models (GLM)
  - Three components
    | A probability distribution from the exponential family
    | A linear predictor $\eta_i = \sum_{k=1}^{p} X_{ik}\beta_k$
    | A link function g such that $E(Y) = \mu = g^{-1}(\eta)$

  - GLM in R

  - binary: logistic regression (predict the probability)
    | Problem: outcomes that have two values. Example: Alive/dead, Win/loss, Success/Failure | Model: $log\frac{P(Y=1|X)}{1-P(Y=1|X)} = W^T X$
    | Odds definiation: $\frac{P(Y=1|X)}{1-P(Y=1|X)}$
    | Logistic Function: $\sigma(a) = \frac{1}{1+e^{-a}}$
    | R example
    | Significance Test: Chisquare Test
    | mathematicalmonk's channel

  - Poisson
    | Problem: counts, rate. Example:calls to a call center, percent of child passing a test
    | Model: $log(Y) = W^T X$
    | R example
    | Log Linear on Titanic

Machine Learning

- General Issues
  - The “Best” Machine Learning Method
    | Interpretable
    | Simple
    | Accurate
    | Fast
    | Scalable
  - prediction study design
    | Split your data into training, test and validation(optional).
    | select model by cross-validation or by experiences
    | extra resource: 1
  - cross validation
    | Method: Hold-out, K-fold, random sampling, leave one out.
    | Purpose: model selection, true error rate estimate
    | extra resource: 1
- ML using R
  - preProcess
    | standarizing
    | scalling
    | log
    | imputing
    | box-cox transform (improve normality)
  - data slicing
  - feature plotting
  - train options
    | Metric: Contious outcome(RMSE, R^2), Categorical outcomes(Accuracy, Kappa)
    | resampling control: method(bootstrapping, boot632, cv, repeatedcv, LOOCV)
  - caret example
    | linear regression
    | multiple covariate ilinear regression
    | tree
    | bagging
    | Random forests
    | boosting
    | Regularized regression
    | combinding predictors
    | time series

Data Product Development

Reporting Tools

- rmarkdown
- knit
- rPresentor
- slidify

UI tools

- shiny
- yhat

- github

R Language Tools

- basics startupjing@github
- packages
  - caret

  - dplyr