1. What is cross-validation? How to do it right?

It's a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. Mainly used in settings where the goal is prediction and one wants to estimate how accurately a model will perform in practice. The goal of cross-validation is to define a data set to test the model in the training phase (i.e. validation data set) in order to limit problems like overfitting, and get an insight on how the model will generalize to an independent data set.

Examples: leave-one-out cross validation, K-fold cross validation

How to do it right?

- the training and validation data sets have to be drawn from the same population
- predicting stock prices: trained for a certain 5-year period, it's unrealistic to treat the subsequent 5-year a draw from the same population
- common mistake: for instance the step of choosing the kernel parameters of a SVM should be cross-validated as well

Bias-variance trade-off for k-fold cross validation:

Leave-one-out cross-validation: gives approximately unbiased estimates of the test error since each training set contains almost the entire data set (n−1observations).
But: we average the outputs of n fitted models, each of which is trained on an almost identical set of observations hence the outputs are highly correlated. Since the variance of a mean of quantities increases when correlation of these quantities increase, the test error estimate from a LOOCV has higher variance than the one obtained with k-fold cross validation

Typically, we choose k=5or k=10, as these values have been shown empirically to yield test error estimates that suffer neither from excessively high bias nor high variance.

2. Is it better to design robust or accurate algorithms?

- The ultimate goal is to design systems with good generalization capacity, that is, systems that correctly identify patterns in data instances not seen before
- The generalization performance of a learning system strongly depends on the complexity of the model assumed
- If the model is too simple, the system can only capture the actual data regularities in a rough manner. In this case, the system has poor generalization properties and is said to suffer from underfitting
- By contrast, when the model is too complex, the system can identify accidental patterns in the training data that need not be present in the test set. These spurious patterns can be the result of random fluctuations or of measurement errors during the data collection process. In this case, the

generalization capacity of the learning system is also poor. The learning system is said to be affected by overfitting

- Spurious patterns, which are only present by accident in the data, tend to have complex forms. This is the idea behind the principle of Occam's razor for avoiding overfitting: simpler models are preferred if more complex models do not significantly improve the quality of the description for the observations
- Quick response: Occam's Razor. It depends on the learning task. Choose the right balance
- Ensemble learning can help balancing bias/variance (several weak learners together = strong learner)

3. How to define/select metrics?

- Type of task: regression? Classification?
- Business goal?
- What is the distribution of the target variable?
- What metric do we optimize for?
- Regression: RMSE (root mean squared error), MAE (mean absolute error), WMAE(weighted mean absolute error), RMSLE (root mean squared logarithmic error)…
- Classification: recall, AUC, accuracy, misclassification error, Cohen's Kappa…

Common metrics in regression:

Mean Squared Error Vs Mean Absolute Error RMSE gives a relatively high weight to large errors. The RMSE is most useful when large errors are particularly undesirable.
The MAE is a linear score: all the individual differences are weighted equally in the average. MAE is more robust to outliers than MSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \overline{y}_i|$$

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(p_i+1) - (log(a_i+1))^2}$$

- Where pi is the ith prediction, ai the ith actual response, log(b) the natural logarithm of bb.
- Weighted Mean Absolute Error
  The weighted average of absolute errors. MAE and RMSE consider that each prediction provides equally precise information about the error variation, i.e. the standard variation of the error term is constant over all the predictions. Examples: recommender systems (differences between past and recent products)

$$WMAE = \frac{1}{\sum wt}\sum_{i=1}^{n} w_i|y_i - \overline{y}_i|$$

Common metrics in classification:

- Recall / Sensitivity / True positive rate:
  High when FN low. Sensitive to unbalanced classes.
  $$Sensitivity = \frac{TP}{TP+FN}$$

- Precision / Positive Predictive Value
  High when FP low. Sensitive to unbalanced classes.
  $$Precision = \frac{TP}{TP+FP}$$

- Specificity / True Negative Rate
  High when FP low. Sensitive to unbalanced classes
  $$Specificity = \frac{TN}{TN+FP}$$

- Accuracy
  High when FP and FN are low. Sensitive to unbalanced classes (see "Accuracy paradox")
-
  $$Accuracy = \frac{TP+TN}{TN+TP+FP+FN}$$

- ROC / AUC
  ROC is a graphical plot that illustrates the performance of a binary classifier (Sensitivity Vs 1− −Specificity or Sensitivity Vs Specificity). They are not sensitive to unbalanced classes.
  AUC is the area under the ROC curve. Perfect classifier: AUC=1, fall on (0,1); 100% sensitivity (no FN) and 100% specificity (no FP)

- Logarithmic loss
  Punishes infinitely the deviation from the true value! It's better to be somewhat wrong than emphatically wrong!
  $$logloss = -\frac{1}{N}\sum_{i=1}^{n}(y_i log(pi) + (1-y_i)log(1-pi))$$

- Misclassification Rate:
  $$Misclassification = \frac{1}{n}\sum_i I(y_i \neq \overline{y}_i)$$

- F1-Score:Used when the target variable is unbalanced.
  $$F1Score = 2 * \frac{precision * Recall}{precision + Recall}$$

4. Explain what regularization is and why it is useful. What are the benefits and drawbacks of specific methods, such as ridge regression and lasso?

- Used to prevent over fitting: improve the generalization of a model
- Decreases complexity of a model
- Introducing a regularization term to a general loss function: adding a term to the minimization problem
- Impose Occam's Razor in the solution

Ridge regression:

- We use an L2 penalty when fitting the model using least squares
- We add to the minimization problem an expression (shrinkage penalty) of the form λ×∑coefficients
- λ: tuning parameter; controls the bias-variance tradeoff; accessed with cross-validation
- A bit faster than the lasso
- The Lasso:

$$\widehat{\beta}^{Ridge} = argmin\,\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta j^2\,\}$$

- We use an L1 penalty when fitting the model using least squares

  Can force regression coefficients to be exactly: feature selection method by itself

$$\beta^{Lasso} = argmin\,\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}||\beta j||\,\}$$

5. Explain what a local optimum is and why it is important in a specific context, such as K-means clustering. What are specific ways of determining if you have a local optimum problem? What can be done to avoid local optima?

A solution that is optimal in within a neighboring set of candidate solutions

In contrast with global optimum: the optimal solution among all others

K-means clustering context:
It's proven that the objective cost function will always decrease until a local optimum is reached.
Results will depend on the initial random cluster assignment

Determining if you have a local optimum problem:
Tendency of premature convergence
Different initialization induces different optima

Avoid local optima in a K-means context: repeat K-means and take the solution that has the lowest cost

6. Assume you need to generate a predictive model using multiple regression. Explain how you intend to validate this model

Validation using R^2:
- % of variance retained by the model

- Issue: $R^2$is always increased when adding variables

- $R^2 = \frac{RSStot - RSSres}{RSStot} = \frac{RSSreg}{RSStot} = 1 - \frac{RSSres}{RSStot}$

Analysis of residuals:
- Heteroskedasticity (relation between the variance of the model errors and the size of an independent variable's observations)
- Scatter plots residuals Vs predictors
- Normality of errors
- Etc. : diagnostic plots

Out-of-sample evaluation: with cross-validation

7. Explain what precision and recall are. How do they relate to the ROC curve?

See question 3. "How to define/select metrics? Do you know compound metrics?".
When using Precision/Recall curves.

8. What is latent semantic indexing? What is it used for? What are the specific limitations of the method?

- Indexing and retrieval method that uses singular value decomposition to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text
- Based on the principle that words that are used in the same contexts tend to have similar meanings
- "Latent": semantic associations between words is present not explicitly but only latently
- For example: two synonyms may never occur in the same passage but should nonetheless have highly associated representations

Used for:

- Learning correct word meanings
- Subject matter comprehension
- Information retrieval
- Sentiment analysis (social network analysis)
  Here's a great tutorial on it.

9. Explain what resampling methods are and why they are useful

- Repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model
- example: repeatedly draw different samples from training data, fit a linear regression to each new sample, and then examine the extent to which the resulting fit differ
- most common are: cross-validation and the bootstrap
- cross-validation: random sampling with no replacement
- bootstrap: random sampling with replacement
- cross-validation: evaluating model performance, model selection (select the appropriate level of flexibility)
- bootstrap: mostly used to quantify the uncertainty associated with a given estimator or statistical learning method

10. What is principal component analysis? Explain the sort of problems you would use PCA for. Also explain its limitations as a method

Statistical method that uses an orthogonal transformation to convert a set of observations of correlated variables into a set of values of linearly uncorrelated variables called principal components.

Reduce the data from n to k dimensions: find the k vectors onto which to project the data so as to minimize the projection error.
Algorithm:
1) Preprocessing (standardization): PCA is sensitive to the relative scaling of the original variable
2) Compute covariance matrix Σ
3) Compute eigenvectors of Σ
4) Choose k principal components so as to retain x% of the variance (typically x=99)
Applications:
1) Compression
- Reduce disk/memory needed to store data
- Speed up learning algorithm. Warning: mapping should be defined only on training set and then applied to test set

2. Visualization: 2 or 3 principal components, so as to summarize data

Limitations:
- PCA is not scale invariant
- The directions with largest variance are assumed to be of most interest
- Only considers orthogonal transformations (rotations) of the original variables
- PCA is only based on the mean vector and covariance matrix. Some distributions (multivariate normal) are characterized by this but some are not
- If the variables are correlated, PCA can achieve dimension reduction. If not, PCA just orders them according to their variances

11. Explain what a false positive and a false negative are. Why is it important these from each other? Provide examples when false positives are more

important than false negatives, false negatives are more important than false positives and when these two types of errors are equally important

- False positive
  Improperly reporting the presence of a condition when it's not in reality. Example: HIV positive test when the patient is actually HIV negative

- False negative
  Improperly reporting the absence of a condition when in reality it's the case. Example: not detecting a disease when the patient has this disease.

When false positives are more important than false negatives:
- In a non-contagious disease, where treatment delay doesn't have any long-term consequences but the treatment itself is grueling
- HIV test: psychological impact

When false negatives are more important than false positives:
- If early treatment is important for good outcomes
- In quality control: a defective item passes through the cracks!
- Software testing: a test to catch a virus has failed

12. What is the difference between supervised learning and unsupervised learning? Give concrete examples

- Supervised learning: inferring a function from labeled training data
- Supervised learning: predictor measurements associated with a response measurement; we wish to fit a model that relates both for better understanding the relation between them (inference) or with the aim to accurately predicting the response for future observations (prediction)
- Supervised learning: support vector machines, neural networks, linear regression, logistic regression, extreme gradient boosting
- Supervised learning examples: predict the price of a house based on the are, size.; churn prediction; predict the relevance of search engine results.
- Unsupervised learning: inferring a function to describe hidden structure of unlabeled data
- Unsupervised learning: we lack a response variable that can supervise our analysis
- Unsupervised learning: clustering, principal component analysis, singular value decomposition; identify group of customers
- Unsupervised learning examples: find customer segments; image segmentation; classify US senators by their voting.

13. What does NLP stand for?

"Natural language processing"!

- Interaction with human (natural) and computers languages
- Involves natural language understanding

Major tasks:
- Machine translation
- Question answering: “what’s the capital of Canada?”
- Sentiment analysis: extract subjective information from a set of documents, identify trends or public opinions in the social media
- Information retrieval

14. What are feature vectors?

- n-dimensional vector of numerical features that represent some object
- term occurrences frequencies, pixels of an image etc.
- Feature space: vector space associated with these vectors

15. When would you use random forests Vs SVM and why?

- In a case of a multi-class classification problem: SVM will require one-against-all method (memory intensive)
- If one needs to know the variable importance (random forests can perform it as well)
- If one needs to get a model fast (SVM is long to tune, need to choose the appropriate kernel and its parameters, for instance sigma and epsilon)
- In a semi-supervised learning context (random forest and dissimilarity measure): SVM can work only in a supervised learning mode

16. How do you take millions of users with 100’s transactions each, amongst 10k’s of products and group the users together in meaningful segments?

1. Some exploratory data analysis (get a first insight)

- Transactions by date
- Count of customers Vs number of items bought
- Total items Vs total basket per customer
- Total items Vs total basket per area

2. Create new features (per customer):

Counts:

- Total baskets (unique days)
- Total items
- Total spent

- Unique product id

Distributions:

- Items per basket
- Spent per basket
- Product id per basket
- Duration between visits
- Product preferences: proportion of items per product cat per basket

3. Too many features, dimension-reduction? PCA?
4. Clustering:

- PCA

5. Interpreting model fit

- View the clustering by principal component axis pairs PC1 Vs PC2, PC2 Vs PC1.
- Interpret each principal component regarding the linear combination it's obtained from; example: PC1=spendy axis (proportion of baskets containing spendy items, raw counts of items and visits)

17. How do you know if one algorithm is better than other?

- In terms of performance on a given data set?
- In terms of performance on several data sets?
- In terms of efficiency?

In terms of performance on several data sets:
- "Does learning algorithm A have a higher chance of producing a better predictor than learning algorithm B in the given context?"
- "Bayesian Comparison of Machine Learning Algorithms on Single and Multiple Datasets", A. Lacoste and F. Laviolette
- "Statistical Comparisons of Classifiers over Multiple Data Sets", Janez Demsar

In terms of performance on a given data set:
- One wants to choose between two learning algorithms
- Need to compare their performances and assess the statistical significance

One approach (Not preferred in the literature):
- Multiple k-fold cross validation: run CV multiple times and take the mean and sd
- You have: algorithm A (mean and sd) and algorithm B (mean and sd)
- Is the difference meaningful? (Paired t-test)

Sign-test (classification context):
Simply counts the number of times A has a better metrics than B and

assumes this comes from a binomial distribution. Then we can obtain a p-value of the Ho test: A and B are equal in terms of performance.
Wilcoxon signed rank test (classification context):
Like the sign-test, but the wins (A is better than B) are weighted and assumed coming from a symmetric distribution around a common median. Then, we obtain a p-value of the Ho test.
Other (without hypothesis testing):
- AUC
- F-Score
- See question 3

18. How do you test whether a new credit risk scoring model works?

- Test on a holdout set
- Kolmogorov-Smirnov test

Kolmogorov-Smirnov test:
- Non-parametric test
- Compare a sample with a reference probability distribution or compare two samples
- Quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution
- Or between the empirical distribution functions of two samples
- Null hypothesis (two-samples test): samples are drawn from the same distribution
- Can be modified as a goodness of fit test
- In our case: cumulative percentages of good, cumulative percentages of bad

19. What is: collaborative filtering, n-grams, cosine distance?

Collaborative filtering:
- Technique used by some recommender systems
- Filtering for information or patterns using techniques involving collaboration of multiple agents: viewpoints, data sources.
1. A user expresses his/her preferences by rating items (movies, CDs.)
2. The system matches this user's ratings against other users' and finds people with most similar tastes
3. With similar users, the system recommends items that the similar users have rated highly but not yet being rated by this user

n-grams:
- Contiguous sequence of n items from a given sequence of text or speech
- "Andrew is a talented data scientist"
- Bi-gram: "Andrew is", "is a", "a talented".
- Tri-grams: "Andrew is a", "is a talented", "a talented data".
- An n-gram model models sequences using statistical properties of n-grams; see: Shannon Game

- More concisely, n-gram model: P(Xi|Xi-(n-1)...Xi-1): Markov model
- N-gram model: each word depends only on the n-1 last words

Issues:
- when facing infrequent n-grams
- solution: smooth the probability distributions by assigning non-zero probabilities to unseen words or n-grams
- Methods: Good-Turing, Backoff, Kneser-Kney smoothing

Cosine distance:
- How similar are two documents?
- Perfect similarity/agreement: 1
- No agreement : 0 (orthogonality)
- Measures the orientation, not magnitude

Given two vectors A and B representing word frequencies:
cosine-similarity(A,B)=$\frac{(\mathbf{A},\mathbf{B})}{||\mathbf{A}||.||\mathbf{B}||}$

20. What is better: good data or good models? And how do you define "good"? Is there a universal good model? Are there any models that are definitely not so good?

- Good data is definitely more important than good models
- If quality of the data wasn't of importance, organizations wouldn't spend so much time cleaning and preprocessing it!
- Even for scientific purpose: good data (reflected by the design of experiments) is very important

How do you define good?
- good data: data relevant regarding the project/task to be handled
- good model: model relevant regarding the project/task
- good model: a model that generalizes on external data sets

Is there a universal good model?
- No, otherwise there wouldn't be the overfitting problem!
- Algorithm can be universal but not the model
- Model built on a specific data set in a specific organization could be ineffective in other data set of the same organization
- Models have to be updated on a somewhat regular basis

Are there any models that are definitely not so good?
- "all models are wrong but some are useful" George E.P. Box
- It depends on what you want: predictive models or explanatory power
- If both are bad: bad model

21. Why is naive Bayes so bad? How would you improve a spam detection algorithm that uses naive Bayes?

- Naïve: the features are assumed independent/uncorrelated

- Assumption not feasible in many cases
- Improvement: decorrelate features (covariance matrix into identity matrix)

22. What are the drawbacks of linear model? Are you familiar with alternatives (Lasso, ridge regression)?

- Assumption of linearity of the errors
- Can't be used for count outcomes, binary outcomes
- Can't vary model flexibility: overfitting problems
- Alternatives: see question 4 about regularization

23. Do you think 50 small decision trees are better than a large one? Why?

- Yes!
- More robust model (ensemble of weak learners that come and make a strong learner)
- Better to improve a model by taking many small steps than fewer large steps
- If one tree is erroneous, it can be auto-corrected by the following
- Less prone to overfitting

24. Why is mean square error a bad measure of model performance? What would you suggest instead?

- see question 3 about metrics in regression
- It puts too much emphasis on large deviations (squared)
- Alternative: mean absolute deviation

25. How can you prove that one improvement you've brought to an algorithm is really an improvement over not doing anything? Are you familiar with A/B testing?

Example with linear regression:
- F-statistic (ANOVA)

$$F = \frac{\frac{RSS1 - RSS2}{p_2 - p_1}}{\frac{RSS_2}{n - p_2}}$$

p1: number of parameters of model 1
p2: number of parameters of model 2
n: number of observations

Under the null hypothesis that model 2 doesn’t provide a significantly better fit than model 1, F will have an F distribution with (p2−p1,n−p2) degrees of freedom. The null hypothesis is rejected if the F calculated from the data is greater than the critical value of the FF distribution for some desired significance level.
Others: AIC/BIC (regression), cross-validation: assessing test error on a test/validation set

26. What do you think about the idea of injecting noise in your data set to test the sensitivity of your models?

- Effect would be similar to regularization: avoid overfitting
- Used to increase robustness

27. Do you know / used data reduction techniques other than PCA? What do you think of step-wise regression? What kind of step-wise techniques are you familiar with?

data reduction techniques other than PCA?:
Partial least squares: like PCR (principal component regression) but chooses the principal components in a supervised way. Gives higher weights to variables that are most strongly related to the response

step-wise regression?
- the choice of predictive variables are carried out using a systematic procedure
- Usually, it takes the form of a sequence of F-tests, t-tests, adjusted R-squared, AIC, BIC
- at any given step, the model is fit using unconstrained least squares
- can get stuck in local optima
- Better: Lasso

step-wise techniques:
- Forward-selection: begin with no variables, adding them when they improve a chosen model comparison criterion
- Backward-selection: begin with all the variables, removing them when it improves a chosen model comparison criterion

Better than reduced data:
Example 1: If all the components have a high variance: which components to discard with a guarantee that there will be no significant loss of the information?
Example 2 (classification):
- One has 2 classes; the within class variance is very high as compared to between class variance
- PCA might discard the very information that separates the two classes

Better than a sample:
- When number of variables is high relative to the number of observations

28. How would you define and measure the predictive power of a metric?

- Predictive power of a metric: the accuracy of a metric's success at predicting the empirical
- They are all domain specific
- Example: in field like manufacturing, failure rates of tools are easily observable. A metric can be trained and the success can be easily measured as the deviation over time from the observed
- In information security: if the metric says that an attack is coming and one should do X. Did the recommendation stop the attack or the attack never happened?

29. Do we always need the intercept term in a regression model?

- It guarantees that the residuals have a zero mean
- It guarantees the least squares slopes estimates are unbiased
- the regression line floats up and down, by adjusting the constant, to a point where the mean of the residuals is zero

30. What are the assumptions required for linear regression? What if some of these assumptions are violated?

1. The data used in fitting the model is representative of the population
2. The true underlying relation between x and y is linear
3. Variance of the residuals is constant (homoscedastic, not heteroscedastic)
4. The residuals are independent
5. The residuals are normally distributed

Predict y from x: 1) + 2)
Estimate the standard error of predictors: 1) + 2) + 3)
Get an unbiased estimation of y from x: 1) + 2) + 3) + 4)
Make probability statements, hypothesis testing involving slope and correlation, confidence intervals: 1) + 2) + 3) + 4) + 5)
Note:
- Common mythology: linear regression doesn't assume anything about the distributions of x and y
- It only makes assumptions about the distribution of the residuals
- And this is only needed for statistical tests to be valid
- Regression can be applied to many purposes, even if the errors are not normally distributed

31. What is collinearity and what to do with it? How to remove multicollinearity?

Collinearity/Multicollinearity:
- In multiple regression: when two or more variables are highly

correlated
- They provide redundant information
- In case of perfect multicollinearity: $\beta = (X^T X)^{-1} X^T y$ doesn't exist, the design matrix isn't invertible
- It doesn't affect the model as a whole, doesn't bias results
- The standard errors of the regression coefficients of the affected variables tend to be large
- The test of hypothesis that the coefficient is equal to zero may lead to a failure to reject a false null hypothesis of no effect of the explanatory (Type II error)
- Leads to overfitting

Remove multicollinearity:
- Drop some of affected variables
- Principal component regression: gives uncorrelated predictors
- Combine the affected variables
- Ridge regression
- Partial least square regression

Detection of multicollinearity:
- Large changes in the individual coefficients when a predictor variable is added or deleted
- Insignificant regression coefficients for the affected predictors but a rejection of the joint hypothesis that those coefficients are all zero (F-test)
- VIF: the ratio of variances of the coefficient when fitting the full model divided by the variance of the coefficient when fitted on its own
- rule of thumb: VIF>5 indicates multicollinearity
- Correlation matrix, but correlation is a bivariate relationship whereas multicollinearity is multivariate

---

32. How to check if the regression model fits the data well?

R squared/Adjusted R squared:

- $R^2 = \frac{RSStot - RSSres}{RSStot} = \frac{RSSreg}{RSStot} = 1 - \frac{RSSres}{RSStot}$

- Describes the percentage of the total variation described by the model
- R2 always increases when adding new variables: adjusted R2 incorporates the model's degrees of freedom

F test:
- Evaluate the hypothesis Ho: all regression coefficients are equal to zero Vs H1: at least one doesn't
- Indicates that R2 is reliable

RMSE:
- Absolute measure of fit (whereas R2 is a relative measure of fit)

33. What is a decision tree?

1. Take the entire data set as input
2. Search for a split that maximizes the "separation" of the classes. A split is any test that divides the data in two (e.g. if variable2>10)
3. Apply the split to the input data (divide step)
4. Re-apply steps 1 to 2 to the divided data
5. Stop when you meet some stopping criteria
6. (Optional) Clean up the tree when you went too far doing splits (called pruning)

Finding a split: methods vary, from greedy search (e.g. C4.5) to randomly selecting attributes and split points (random forests)

Purity measure: information gain, Gini coefficient, Chi Squared values

Stopping criteria: methods vary from minimum size, particular confidence in prediction, purity criteria threshold

Pruning: reduced error pruning, out of bag error pruning (ensemble methods)

34. What impurity measures do you know?

Gini

$$Gini = 1 - \sum_j pj^2$$

Information Gain/Deviance

$$InformationGain = \sum_j p_j log_2 p_j$$

Better than Gini when P_j are very small: multiplying very small numbers leads to rounding errors, we can instead take logs.

35. What is random forest? Why is it good?

Random forest? (Intuition):
- Underlying principle: several weak learners combined provide a strong learner
- Builds several decision trees on bootstrapped training samples of data
- On each tree, each time a split is considered, a random sample of mm predictors is chosen as split candidates, out of all pp predictors
- Rule of thumb: at each split m=$\sqrt{p}$
- Predictions: at the majority rule

Why is it good?
- Very good performance (decorrelates the features)
- Can model non-linear class boundaries

- Generalization error for free: no cross-validation needed, gives an unbiased estimate of the generalization error as the trees is built
- Generates variable importance

36. How do we train a logistic regression model? How do we interpret its coefficients?

$log(odds) = log(\frac{p(y=1|x)}{p(y=0|x)})$ = is a linear function of the input features

Minimization objective/Cost function:

$$J(\beta) = -\frac{1}{m}\sum_{i=1}^{n} y^i \log(h_\beta(x^i)) + 1 - y^i \log(1 - h_\beta(x^i))$$

Where: $h\beta(x) = g(\beta^T x)$ and $g(z) = \frac{1}{1+e^{-z}}$ (sigmoid function)

Intuition:

- if $y_i = 0, J(\beta) = \log(1 - h_\beta(x)_i)$ will converge to ∞ as $h_\beta(x)_i$ becomes far from 0
- Converse: when $y_i = 1, J(\beta) = \log(h_\beta(x)_i)$, will converge to ∞ as $h_\beta(x)_i$ becomes far from 1

Interpretation of the coefficients: the increase of log odds for the increase of one unit of a predictor, given all the other predictors are fixed.

37. What is the maximal margin classifier? How this margin can be achieved?

- When the data can be perfectly separated using a hyperplane, there actually exists an infinite number of these hyperplanes
- Intuition: a hyperplane can usually be shifted a tiny bit up, or down, or rotated, without coming into contact with any of the observations
- Large margin classifier: choosing the hyperplance that is farthest from the training observations
- This margin can be achievec using support vectors

38. What is a kernel? Explain the kernel trick

39. Which kernels do you know? How to choose a kernel?

- Gaussian kernel
- Linear kernel
- Polynomial kernel
- Laplace kernel
- Esoteric kernels: string kernels, chi-square kernels

- If number of features is large (relative to number of observations): SVM with linear kernel ; e.g. text classification with lots of words, small training example
- If number of features is small, number of observations is intermediate: Gaussian kernel
- If number of features is small, number of observations is small: linear kernel

40. Is it beneficial to perform dimensionality reduction before fitting an SVM? Why or why not?

- When the number of features is large comparing to the number of observations (e.g. document-term matrix)
- SVM will perform better in this reduced space

41. What is an Artificial Neural Network? What is back propagation?

42. What is curse of dimensionality? How does it affect distance and similarity measures?

- Refers to various phenomena that arise when analyzing and organizing data in high dimensional spaces
- Common theme: when number of dimensions increases, the volume of the space increases so fast that the available data becomes sparse
- Issue with any method that requires statistical significance: the amount of data needed to support the result grows exponentially with the dimensionality
- Issue when algorithms don't scale well on high dimensions typically when $o(n^{kn})$
- Everything becomes far and difficult to organize

Illustrative example: compare the proportion of an inscribed hypersphere with radius rr and dimension d to that of a hypercube with edges of length 2r

- Volume of such a sphere is $Vsphere = \frac{2r^d \pi^{\frac{d}{2}}}{d\tau\left(\frac{d}{2}\right)}$

- The volume of the cube is: $Vcube = 2r^d$

As d increases (space dimension), the volume of hypersphere becomes insignificant relative to the volume of the hypercube:

$$lim_{d \to \infty} \frac{V_{sphere}}{V_{cube}} = \frac{\pi^{\frac{d}{2}}}{d_2^{d-1} \square\left(\frac{\square}{\square}\right)} = 0$$

- Nearly all of the dimensional space is far away from the center
- It consists almost entirely of the corners of the hypercube, with no middle!

43. What is Ax=b? How to solve it?

- A matrix equation/a system of linear equations
- calculate the inverse of AA (if non singular)
- can be done using Gaussian elimination

44. How do we multiply matrices?

- A$\in R^{n\times m}$ *and* $\mathbf{B} \in R^{m\times p}$
- Each entry: ABij=$\sum_{k=1}^{m} A_{ik}B_{kj}$

45. What is singular value decomposition? What is an eigenvalue? And what is an eigenvector?

46. What's the relationship between PCA and SVD?

47. Can you derive the ordinary least square regression formula?

48. What is the difference between a convex function and non-convex?

49. What is gradient descent method? Will gradient descent methods always converge to the same point?

50. What the Newton's method is?

51. Imagine you have N pieces of rope in a bucket. You reach in and grab one end-piece, then reach in and grab another end-piece, and tie those two together. What is the expected value of the number of loops in the bucket?

- There are n entirely unattached pieces of rope in a bucket
- A loop: any number of rope attached in a closed chain
- Suppose the expected number of loops for n−1 pieces of rope is denoted Ln−1
- Consider the bucket of n pieces of rope; there are 2n rope ends

Pick an end of rope. Of the remaining 2n−1 ends of rope, only one end creates a loop (the other end of the same piece of rope). There are then n−1 untied pieces of rope. The rest of the time, two separates pieces of rope are tied together and there are effectively n−1 untied pieces of rope. The recurrence is therefore:

- $Ln=\frac{1}{2n-1}+L_{n-1}$

Clearly, L1=1 so:

- $Ln=\sum_{k=1}^{n}\frac{1}{2k-1}=H_{2n}-\frac{H_n}{2}$

- Where Hk is the kth harmonic number
  Since $Hk = \gamma + \ln K$ k for large-ish k, where gamma=0.57722. is the Euler-Mascheroni constant, we have:
- $Ln = \ln(2n) - \frac{\ln(n)}{2} = \ln 2\sqrt{n}$

Thanks to Brian Tung.

Statistics

1. How do you assess the statistical significance of an insight?

- is this insight just observed by chance or is it a real insight?

Statistical significance can be accessed using hypothesis testing:
- Stating a null hypothesis which is usually the opposite of what we wish to test (classifiers A and B perform equivalently, Treatment A is equal of treatment B)
- Then, we choose a suitable statistical test and statistics used to reject the null hypothesis
- Also, we choose a critical region for the statistics to lie in that is extreme enough for the null hypothesis to be rejected (p-value)
- We calculate the observed test statistics from the data and check whether it lies in the critical region

Common tests:
- One sample Z test
- Two-sample Z test
- One sample t-test
- paired t-test
- Two sample pooled equal variances t-test
- Two sample unpooled unequal variances t-test and unequal sample sizes (Welch's t-test)
- Chi-squared test for variances
- Chi-squared test for goodness of fit
- Anova (for instance: are the two regression models equals? F-test)
- Regression F-test (i.e: is at least one of the predictor useful in predicting the response?)

2. Explain what a long-tailed distribution is and provide three examples of relevant phenomena that have long tails. Why are they important in classification and regression problems?

- In long tailed distributions, a high frequency population is followed by a low frequency population, which gradually tails off asymptotically
- Rule of thumb: majority of occurrences (more than half, and when Pareto principles applies, 80%) are accounted for by the first 20% items in the distribution
- The least frequently occurring 80% of items are more important as a proportion of the total population
- Zipf's law, Pareto distribution, power laws

Examples:
1) Natural language
- Given some corpus of natural language - The frequency of any word is inversely proportional to its rank in the frequency table
- The most frequent word will occur twice as often as the second most frequent, three times as often as the third most frequent…
- "The" accounts for 7% of all word occurrences (70000 over 1 million)
- "of" accounts for 3.5%, followed by "and"…
- Only 135 vocabulary items are needed to account for half the English corpus!

2. Allocation of wealth among individuals: the larger portion of the wealth of any society is controlled by a smaller percentage of the people

3. File size distribution of Internet Traffic

Additional: Hard disk error rates, values of oil reserves in a field (a few large fields, many small ones), sizes of sand particles, sizes of meteorites

Importance in classification and regression problems:
- Skewed distribution
- Which metrics to use? Accuracy paradox (classification), F-score, AUC
- Issue when using models that make assumptions on the linearity (linear regression): need to apply a monotone transformation on the data (logarithm, square root, sigmoid function…)
- Issue when sampling: your data becomes even more unbalanced! Using of stratified sampling of random sampling, SMOTE ("Synthetic Minority Over-sampling Technique", NV Chawla) or anomaly detection approach

3. What is the Central Limit Theorem? Explain it. Why is it important?

The CLT states that the arithmetic mean of a sufficiently large number of iterates of independent random variables will be approximately normally distributed regardless of the underlying distribution. i.e: the sampling distribution of the sample mean is normally distributed.
- Used in hypothesis testing
- Used for confidence intervals
- Random variables must be iid: independent and identically

distributed
- Finite variance

4. What is statistical power?

- sensitivity of a binary hypothesis test
- Probability that the test correctly rejects the null hypothesis H0H0 when the alternative is true H1
- Ability of a test to detect an effect, if the effect actually exists
- Power=P(reject H0|H1 is true)Power
- As power increases, chances of Type II error (false negative) decrease
- Used in the design of experiments, to calculate the minimum sample size required so that one can reasonably detects an effect. i.e: "how many times do I need to flip a coin to conclude it is biased?"
- Used to compare tests. Example: between a parametric and a non-parametric test of the same hypothesis

---

5. Explain selection bias (with regard to a dataset, not variable selection). Why is it important? How can data management procedures such as missing data handling make it worse?

- Selection of individuals, groups or data for analysis in such a way that proper randomization is not achieved

Types:
- Sampling bias: systematic error due to a non-random sample of a population causing some members to be less likely to be included than others
- Time interval: a trial may terminated early at an extreme value (ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all the variables have similar means
- Data: "cherry picking", when specific subsets of the data are chosen to support a conclusion (citing examples of plane crashes as evidence of airline flight being unsafe, while the far more common example of flights that complete safely)
- Studies: performing experiments and reporting only the most favorable results
- Can lead to unaccurate or even erroneous conclusions
- Statistical methods can generally not overcome it

Why data handling make it worse?
- Example: individuals who know or suspect that they are HIV positive are less likely to participate in HIV surveys
- Missing data handling will increase this effect as it's based on most HIV negative
-Prevalence estimates will be unaccurate

6. Provide a simple example of how an experimental design can help answer a question about behavior. How does experimental data contrast with observational data?

- You are researching the effect of music-listening on studying efficiency
- You might divide your subjects into two groups: one would listen to music and the other (control group) wouldn't listen anything!
- You give them a test
- Then, you compare grades between the two groups

Differences between observational and experimental data:
- Observational data: measures the characteristics of a population by studying individuals in a sample, but doesn't attempt to manipulate or influence the variables of interest
- Experimental data: applies a treatment to individuals and attempts to isolate the effects of the treatment on a response variable

Observational data: find 100 women age 30 of which 50 have been smoking a pack a day for 10 years while the other have been smoke free for 10 years. Measure lung capacity for each of the 100 women. Analyze, interpret and draw conclusions from data.

Experimental data: find 100 women age 20 who don't currently smoke. Randomly assign 50 of the 100 women to the smoking treatment and the other 50 to the no smoking treatment. Those in the smoking group smoke a pack a day for 10 years while those in the control group remain smoke free for 10 years. Measure lung capacity for each of the 100 women.
Analyze, interpret and draw conclusions from data.

7. Is mean imputation of missing data acceptable practice? Why or why not?

- Bad practice in general
- If just estimating means: mean imputation preserves the mean of the observed data
- Leads to an underestimate of the standard deviation
- Distorts relationships between variables by "pulling" estimates of the correlation toward zero

---

8. What is an outlier? Explain how you might screen for outliers and what would you do if you found them in your dataset. Also, explain what an inlier is and how you might screen for them and what would you do if you found them in your dataset

Outliers:
- An observation point that is distant from other observations

- Can occur by chance in any distribution
- Often, they indicate measurement error or a heavy-tailed distribution
- Measurement error: discard them or use robust statistics
- Heavy-tailed distribution: high skewness, can't use tools assuming a normal distribution
- Three-sigma rules (normally distributed data): 1 in 22 observations will differ by twice the standard deviation from the mean
- Three-sigma rules: 1 in 370 observations will differ by three times the standard deviation from the mean

Three-sigma rules example: in a sample of 1000 observations, the presence of up to 5 observations deviating from the mean by more than three times the standard deviation is within the range of what can be expected, being less than twice the expected number and hence within 1 standard deviation of the expected number (Poisson distribution).

If the nature of the distribution is known a priori, it is possible to see if the number of outliers deviate significantly from what can be expected. For a given cutoff (samples fall beyond the cutoff with probability p), the number of outliers can be approximated with a Poisson distribution with lambda=pn. Example: if one takes a normal distribution with a cutoff 3 standard deviations from the mean, p=0.3% and thus we can approximate the number of samples whose deviation exceed 3 sigmas by a Poisson with lambda=3

Identifying outliers:
- No rigid mathematical method
- Subjective exercise: be careful
- Boxplots
- QQ plots (sample quantiles Vs theoretical quantiles)

Handling outliers:
- Depends on the cause
- Retention: when the underlying model is confidently known
- Regression problems: only exclude points which exhibit a large degree of influence on the estimated coefficients (Cook's distance)

Inlier:
- Observation lying within the general distribution of other observed values
- Doesn't perturb the results but are non-conforming and unusual
- Simple example: observation recorded in the wrong unit (°F instead of °C)

Identifying inliers:
- Mahalanobi's distance
- Used to calculate the distance between two random vectors
- Difference with Euclidean distance: accounts for correlations
- Discard them

9. How do you handle missing data? What imputation techniques do you recommend?

- If data missing at random: deletion has no bias effect, but decreases the power of the analysis by decreasing the effective sample size
- Recommended: Knn imputation, Gaussian mixture imputation

10. You have data on the durations of calls to a call center. Generate a plan for how you would code and analyze these data. Explain a plausible scenario for what the distribution of these durations might look like. How could you test, even graphically, whether your expectations are borne out?

1. Exploratory data analysis

- Histogram of durations
- histogram of durations per service type, per day of week, per hours of day (durations can be systematically longer from 10am to 1pm for instance), per employee…

2. Distribution: lognormal?
3. Test graphically with QQ plot: sample quantiles of log(durations)log(durations) Vs normal quantiles

11. Explain likely differences between administrative datasets and datasets gathered from experimental studies. What are likely problems encountered with administrative data? How do experimental methods help alleviate these problems? What problem do they bring?

Advantages:
- Cost
- Large coverage of population
- Captures individuals who may not respond to surveys
- Regularly updated, allow consistent time-series to be built-up

Disadvantages:
- Restricted to data collected for administrative purposes (limited to administrative definitions. For instance: incomes of a married couple, not individuals, which can be more useful)
- Lack of researcher control over content
- Missing or erroneous entries
- Quality issues (addresses may not be updated or a postal code is provided only)
- Data privacy issues
- Underdeveloped theories and methods (sampling methods…)

12. You are compiling a report for user content uploaded every month and notice a spike in uploads in October. In particular, a spike in picture uploads. What might you think is the cause of this, and how would you test it?

- Halloween pictures?
- Look at uploads in countries that don’t observe Halloween as a sort of counter-factual analysis
- Compare uploads mean in October and uploads means with September: hypothesis testing

13. You’re about to get on a plane to Seattle. You want to know if you should bring an umbrella. You call 3 random friends of yours who live there and ask each independently if it’s raining. Each of your friends has a 2/3 chance of telling you the truth and a 1/3 chance of messing with you by lying. All 3 friends tell you that “Yes” it is raining. What is the probability that it’s actually raining in Seattle?

- All say yes: all three lie or three say the truth
- P("all say the truth")=$\frac{2}{3}^3=\frac{8}{27}$
- P("all lie")=$\frac{1}{3}^3=\frac{1}{27}$
- P("all yes")=$\frac{1}{27}+\frac{8}{27}=\frac{1}{3}$
- Out of these numbers, there is $\frac{\frac{8}{27}}{\frac{1}{8}}=\frac{8}{9}$ chance it’s actually raining

14. There’s one box - has 12 black and 12 red cards, 2nd box has 24 black and 24 red; if you want to draw 2 cards at random from one of the 2 boxes, which box has the higher probability of getting the same color? Can you tell intuitively why the 2nd box has a higher probability

- First select: for both, then and ; compare them
- $\frac{B}{A}=\frac{529}{517}$

15. What is: lift, KPI, robustness, model fitting, design of experiments, 80/20 rule?

Lift:
It’s measure of performance of a targeting model (or a rule) at predicting or classifying cases as having an enhanced response (with respect to the population as a whole), measured against a random choice targeting model. Lift is simply: target response/average response.

Suppose a population has an average response rate of 5% (mailing for instance). A certain model (or rule) has identified a segment with a response rate of 20%, then lift=20/5=4

Typically, the modeler seeks to divide the population into quantiles, and rank the quantiles by lift. He can then consider each

quantile, and by weighing the predicted response rate against the cost, he can decide to market that quantile or not.
“if we use the probability scores on customers, we can get 60% of the total responders we’d get mailing randomly by only mailing the top 30% of the scored customers”.

KPI:
- Key performance indicator
- A type of performance measurement
- Examples: 0 defects, 10/10 customer satisfaction
- Relies upon a good understanding of what is important to the organization

More examples:

Marketing & Sales:
- New customers acquisition
- Customer attrition
- Revenue (turnover) generated by segments of the customer population
- Often done with a data management platform

IT operations:
- Mean time between failure
- Mean time to repair

Robustness:
- Statistics with good performance even if the underlying distribution is not normal
- Statistics that are not affected by outliers
- A learning algorithm that can reduce the chance of fitting noise is called robust
- Median is a robust measure of central tendency, while mean is not
- Median absolute deviation is also more robust than the standard deviation

Model fitting:
- How well a statistical model fits a set of observations
- Examples: AIC, R2, Kolmogorov-Smirnov test, Chi 2, deviance (glm)

Design of experiments:
The design of any task that aims to describe or explain the variation of information under conditions that are hypothesized to reflect the variation.
In its simplest form, an experiment aims at predicting the outcome by changing the preconditions, the predictors.
- Selection of the suitable predictors and outcomes
- Delivery of the experiment under statistically optimal conditions
- Randomization
- Blocking: an experiment may be conducted with the same equipment to avoid any unwanted variations in the input
- Replication: performing the same combination run more than once, in order to get an estimate for the amount of random error that could be part of the process
- Interaction: when an experiment has 3 or more variables, the situation in which the interaction of two variables on a third is not additive

80/20 rule:
- Pareto principle

- 80% of the effects come from 20% of the causes
- 80% of your sales come from 20% of your clients
- 80% of a company complaints come from 20% of its customers

16. Define: quality assurance, six sigma.

Quality assurance:
- A way of preventing mistakes or defects in manufacturing products or when delivering services to customers
- In a machine learning context: anomaly detection

Six sigma:
- Set of techniques and tools for process improvement
- 99.99966% of products are defect-free products (3.4 per 1 million)
- 6 standard deviation from the process mean

17. Give examples of data that does not have a Gaussian distribution, nor log-normal.

- Allocation of wealth among individuals
- Values of oil reserves among oil fields (many small ones, a small number of large ones)

18. What is root cause analysis? How to identify a cause vs. a correlation? Give examples

Root cause analysis:
- Method of problem solving used for identifying the root causes or faults of a problem
- A factor is considered a root cause if removal of it prevents the final undesirable event from recurring

Identify a cause vs. a correlation:
- Correlation: statistical measure that describes the size and direction of a relationship between two or more variables. A correlation between two variables doesn’t imply that the change in one variable is the cause of the change in the values of the other variable
- Causation: indicates that one event is the result of the occurrence of the other event; there is a causal relationship between the two events
- Differences between the two types of relationships are easy to identify, but establishing a cause and effect is difficult

Example: sleeping with one’s shoes on is strongly correlated with waking up with a headache. Correlation-implies-causation fallacy: therefore, sleeping with one’s shoes causes headache.
More plausible explanation: both are caused by a third factor: going to bed drunk.

Identify a cause Vs a correlation: use of a controlled study
- In medical research, one group may receive a placebo (control) while the other receives a treatment If the two groups have noticeably different outcomes, the different experiences may have caused the different outcomes

19. Give an example where the median is a better measure than the mean

When data is skewed

20. Given two fair dices, what is the probability of getting scores that sum to 4? to 8?

- Total: 36 combinations
- Of these, 3 involve a score of 4: (1,3), (3,1), (2,2)
- So: 3/36=1/12
- Considering a score of 8: (2,6), (3,5), (4,4), (6,2), (5,3)
- So: 5/36

21. What is the Law of Large Numbers?

- A theorem that describes the result of performing the same experiment a large number of times
- Forms the basis of frequency-style thinking
- It says that the sample mean, the sample variance and the sample standard deviation converge to what they are trying to estimate
- Example: roll a dice, expected value is 3.5. For a large number of experiments, the average converges to 3.5

22. How do you calculate needed sample size?

Estimate a population mean:
- General formula is $ME = t \times \frac{s}{\sqrt{n}}$ or $ME = z \times \frac{s}{\sqrt{n}}$

- ME is the desired margin of error
- t is the t score or z score that we need to use to calculate our confidence interval
- s is the standard deviation

Example: we would like to start a study to estimate the average internet usage of households in one week for our business plan. How many households must we randomly select to be 95% sure that the sample mean is within 1minute from the true mean of the population? A previous survey of household usage has shown a standard deviation of 6.95 minutes.

- Z score corresponding to a 95% interval: 1.96 (97.5%, $\frac{\alpha}{2} = 0.025$
- s=6.95
- $n = \left(\frac{z \times s}{ME}\right)^2 = (1.96 \times 6.95)^2 = 13.62^2 = 186$
- *Estimate a proportion:*

  - Similar: $ME = Z \times \sqrt{\frac{p(1-P)}{n}}$

- Example: a professor in Harvard wants to determine the proportion of students who support gay marriage. She asks "how large a sample do I need?"
  She wants a margin of error of less than 2.5%, she has found a previous survey which indicates a proportion of 30%.
  $n = 0.3 \times \frac{0.7}{0.025^2}$

23. When you sample, what bias are you inflicting?

Selection bias:
- An online survey about computer use is likely to attract people more interested in technology than in typical

Under coverage bias:
- Sample too few observations from a segment of population

Survivorship bias:
- Observations at the end of the study are a non-random set of those present at the beginning of the investigation
- In finance and economics: the tendency for failed companies to be excluded from performance studies because they no longer exist

24. How do you control for biases?

- Choose a representative sample, preferably by a random method
- Choose an adequate size of sample
- Identify all confounding factors if possible
- Identify sources of bias and include them as additional predictors in statistical analyses
- Use randomization: by randomly recruiting or assigning subjects in a study, all our experimental groups have an equal chance of being influenced by the same bias

Notes:
- Randomization: in randomized control trials, research participants are assigned by chance, rather than by choice to either the experimental group or the control group.
- Random sampling: obtaining data that is representative of the population of interest

25. What are confounding variables?

- Extraneous variable in a statistical model that correlates directly or inversely with both the dependent and the independent variable
- A spurious relationship is a perceived relationship between an independent variable and a dependent variable that has been estimated incorrectly
- The estimate fails to account for the confounding factor
- See Question 18 about root cause analysis

26. What is A/B testing?

- Two-sample hypothesis testing
- Randomized experiments with two variants: A and B
- A: control; B: variation
- User-experience design: identify changes to web pages that increase clicks on a banner
- Current website: control; NULL hypothesis
- New version: variation; alternative hypothesis

27. An HIV test has a sensitivity of 99.7% and a specificity of 98.5%. A subject from a population of prevalence 0.1% receives a positive test result. What is the precision of the test (i.e the probability he is HIV positive)?

Bayes rule: $P(Actu+|Pred+) = \frac{p(Pred_+|Actu_+)\times p(Actu_+)}{p(Pred_+|Actu_+)+p(Pred_+|Actu_-)p(Actu_-)}$

We have: $\frac{Sensitivity \times prevalence}{Sensitivity \times Prevalence + (1-Specificity)\times(1-Prevalence)} = \frac{0.997\times 0.001}{0.997\times 0.001+0.15\times 0.999} = 0.62$

28. Infection rates at a hospital above a 1 infection per 100 person days at risk are considered high. An hospital had 10 infections over the last 1787 person days at risk. Give the p-value of the correct one-sided test of whether the hospital is below the standard

One-sided test, assume a Poisson distribution
Ho: lambda=0.01 ; H1:lambda>0.01
R code:

```
ppois(10,1787*0.01)
## [1] 0.03237153
```

29. You roll a biased coin (p(head)=0.8) five times. What's the probability of getting three or more heads?

- 5 trials, p=0.8
  P("3 or more heads")= $\binom{3}{5} \times 0.8^3 \times 0.8 \times 0.2^2 + \binom{4}{1} \times 0.8^4 \times 0.2^1 + \binom{5}{5} * 0.8^5 * 0.2^0 = 0.94$

30. A random variable X is normal with mean 1020 and standard deviation 50. Calculate P(X>1200)

X\~N(1020,50)   Our new quantile: z $\frac{1200-1020}{50} = 3.6$

R Code:

```
pnorm(3.6,lower.tail=F)
## [1] 0.0001591086
```

31. Consider the number of people that show up at a bus station is Poisson with mean 2.5/h. What is the probability that at most three people show up in a four hour period?

X\~Poisson(λ=2.5×t)
R code:

```
ppois(3,lambda=2.5*4)
## [1] 0.01033605
```

32. You are running for office and your pollster polled hundred people. 56 of them claimed they will vote for you. Can you relax?

Quick:

- Intervals take the form $p \pm z \times \sqrt{\frac{1}{n} \times p \times (1-p)}$

- We know that p(1-p) is maximized at $\frac{1}{2}$ and z=1.96 is the relevant quantile for a 95% confidence interval

- So: p± $\frac{1}{\sqrt{n}}$ is a quick estimate for p

- Here: $1/\sqrt{100} = 0.1$ so 95% of the intervals would be [46,66]

- It's not enough!

33. Geiger counter records 100 radioactive decays in 5 minutes. Find an approximate 95% interval for the number of decays per hour.

- Start by finding a 95% interval for radioactive decay in a 5 minutes period
- The estimated standard deviation is $\sqrt{100} = 10$
- 1the interval is $\hat{\lambda} \pm 1.96 \times 10 = 100 \pm 19.6$
- So, per hour: [964.8,1435.2]

34. The homicide rate in Scotland fell last year to 99 from 115 the year before. Is this reported change really networthy?

- Consider the homicides as independent; a Poisson distribution can be a reasonable model
- 95% interval for the true homicide rate is $115 \pm 2 = \sqrt{115} = 115 \pm 22 =$[94,137]
- It's not reasonable to conclude that there has been a reduction in the true rate

35. Consider influenza epidemics for two parent heterosexual families. Suppose that the probability is 17% that at least one of the parents has contracted the disease. The probability that the father has contracted influenza is 12% while the probability that both the mother and father have contracted the disease is 6%. What is the probability that the mother has contracted influenza?

- P("MotherorFather")=P("Mother")+P("Father")−P("MotherandFather") Hence: P("Mother")=0.17+0.06−0.12=0.11

36. Suppose that diastolic blood pressures (DBPs) for men aged 35-44 are normally distributed with a mean of 80 (mm Hg) and a standard deviation of 10. About what is the probability that a random 35-44 year old has a DBP less than 70?

- One standard deviation below the mean: $\frac{32}{2} = 16$

37. In a population of interest, a sample of 9 men yielded a sample average brain volume of 1,100cc and a standard deviation of 30cc. What is a 95% Student's T confidence interval for the mean brain volume in this new population?

- Standard error of the mean: $\frac{30}{\sqrt{9}} = 10$
- Relevant t quantile: 97.5%
  R code:

```
1100+c(-1,1)*qt(0.975,df=9-1)*10
## [1] 1076.94 1123.06
```

38. A diet pill is given to 9 subjects over six weeks. The average difference in weight (follow up - baseline) is -2 pounds. What would the standard deviation of the difference in weight have to be for the upper endpoint of the 95% T confidence interval to touch 0?

- find σ=$2 \times (\frac{\sqrt{9}}{t_{97.5}})$
- R code:

```
2*3/qt(0.975,df=8)
## [1] 2.601903
```

39. In a study of emergency room waiting times, investigators consider a new and the standard triage systems. To test the systems, administrators selected 20 nights and randomly assigned the new triage system to be used on 10 nights and the standard system on the remaining 10 nights. They calculated the nightly median waiting time (MWT) to see a physician. The average MWT for the new system was 3 hours with a variance of 0.60 while the average MWT for the old system was 5 hours with a variance of 0.68. Consider the 95% confidence interval estimate for the differences of the mean MWT associated with the new system. Assume a constant variance. What is the interval? Subtract in this order (New System - Old System).

t confidence interval for the difference of the means assuming equal variances:

(new-old)$\pm \grave{t} \times sp \times \sqrt{\left(\frac{1}{n_1}\right)^2 + \left(\frac{1}{n_2}\right)^2}$

t’: 97.5% quantile, with 20+10-2=28 degrees of freedom: 2.1

- sp: pooled variance, $\sqrt{\frac{(0.6^2 \times 9 + 0.68^2 \times 9)}{(10+10-2)}} = 0.8$
- $\sqrt{\frac{1}{10} + \frac{1}{10}} = 0.44$
- We get [-2.75,-1.25]

40. To further test the hospital triage system, administrators selected 200 nights and randomly assigned a new triage system to be used on 100 nights and a

standard system on the remaining 100 nights. They calculated the nightly median waiting time (MWT) to see a physician. The average MWT for the new system was 4 hours with a standard deviation of 0.5 hours while the average MWT for the old system was 6 hours with a standard deviation of 2 hours. Consider the hypothesis of a decrease in the mean MWT associated with the new treatment. What does the 95% independent group confidence interval with unequal variances suggest vis a vis this hypothesis? (Because there's so many observations per group, just use the Z quantile instead of the T.)

- Z confidence interval for the differences of the means assuming unequal variances: (new-old)$\pm z' \times sp \times \left(\frac{1}{n1}\right)^2 + \left(\frac{1}{n2}\right)^2$
- Z97.5 quantile
- sp: pooled variance, $\sqrt{\frac{0.5^2 \times 99 + 2^2 \times 99}{100+100-2}}$ =1.458

  we get: [1.6,2.4]

Process & Miscellaneous

1. How to optimize algorithms? (parallel processing and/or faster algorithms). Provide examples for both

"Premature optimization is the root of all evil"; Donald Knuth

Parallel processing: for instance in R with a single machine.
- doParallel and foreach package
- doParallel: parallel backend, will select n-cores of the machine
- for each: assign tasks for each core
- using Hadoop on a single node
- using Hadoop on multi-node

Faster algorithm:
- In computer science: Pareto principle; 90% of the execution time is spent executing 10% of the code
- Data structure: affect performance
- Caching: avoid unnecessary work
- Improve source code level

For instance: on early C compilers, WHILE(something) was slower than FOR(;;), because WHILE evaluated "something" and then had a conditional jump which tested if it was true while FOR had unconditional jump.

2. Examples of NoSQL architecture

- Key-value: in a key-value NoSQL database, all of the data within consists of an indexed key and a value. Cassandra, DynamoDB
- Column-based: designed for storing data tables as sections of columns of data rather than as rows of data. HBase, SAP HANA

- Document Database: map a key to some document that contains structured information. The key is used to retrieve the document. MongoDB, CouchDB
- Graph Database: designed for data whose relations are well-represented as a graph and has elements which are interconnected, with an undetermined number of relations between them. Polyglot Neo4J

3. Provide examples of machine-to-machine communications

Telemedicine
- Heart patients wear specialized monitor which gather information regarding heart state
- The collected data is sent to an electronic implanted device which sends back electric shocks to the patient for correcting incorrect rhythms

Product restocking
- Vending machines are capable of messaging the distributor whenever an item is running out of stock

4. Compare R and Python

R
- Focuses on better, user friendly data analysis, statistics and graphical models
- The closer you are to statistics, data science and research, the more you might prefer R
- Statistical models can be written with only a few lines in R
- The same piece of functionality can be written in several ways in R
- Mainly used for standalone computing or analysis on individual servers
- Large number of packages, for anything!

Python
- Used by programmers that want to delve into data science
- The closer you are working in an engineering environment, the more you might prefer Python
- Coding and debugging is easier mainly because of the nice syntax
- Any piece of functionality is always written the same way in Python
- When data analysis needs to be implemented with web apps
- Good tool to implement algorithms for production use

5. Is it better to have 100 small hash tables or one big hash table, in memory, in terms of access speed (assuming both fit within RAM)? What do you think about in-database analytics?

Hash tables:
- Average case O(1)O(1) lookup time
- Lookup time doesn't depend on size
Even in terms of memory:
- O(n)O(n) memory
- Space scales linearly with number of elements
- Lots of dictionaries won't take up significantly less space than a larger one

In-database analytics:
- Integration of data analytics in data warehousing functionality
- Much faster and corporate information is more secure, it doesn't leave the enterprise data warehouse

Good for real-time analytics: fraud detection, credit scoring, transaction processing, pricing and margin analysis, behavioral ad targeting and recommendation engines

6. What is star schema? Lookup tables?

The star schema is a traditional database schema with a central (fact) table (the "observations", with database "keys" for joining with satellite tables, and with several fields encoded as ID's). Satellite tables map ID's to physical name or description and can be "joined" to the central fact table using the ID fields; these tables are known as lookup tables, and are particularly useful in real-time applications, as they save a lot of memory. Sometimes star schemas involve multiple layers of summarization (summary tables, from granular to less granular) to retrieve information faster.

Lookup tables:
- Array that replace runtime computations with a simpler array indexing operation

7. What is the life cycle of a data science project ?

1. Data acquisition
   Acquiring data from both internal and external sources, including social media or web scraping. In a steady state, data extraction and routines should be in place, and new sources, once identified would be acquired following the established processes

2. Data preparation
   Also called data wrangling: cleaning the data and shaping it into a suitable form for later analyses. Involves exploratory data analysis and feature extraction.

3. Hypothesis & modelling
   Like in data mining but not with samples, with all the data instead. Applying machine learning techniques to all the data. A key sub-step: model selection. This involves preparing a training set for model candidates, and validation and test sets for comparing model performances, selecting the best

performing model, gauging model accuracy and preventing overfitting

4. Evaluation & interpretation

Steps 2 to 4 are repeated a number of times as needed; as the understanding of data and business becomes clearer and results from initial models and hypotheses are evaluated, further tweaks are performed. These may sometimes include step5 and be performed in a pre-production.

5. Deployment
6. Operations
   Regular maintenance and operations. Includes performance tests to measure model performance, and can alert when performance goes beyond a certain acceptable threshold
7. Optimization
   Can be triggered by failing performance, or due to the need to add new data sources and retraining the model or even to deploy new versions of an improved model

Note: with increasing maturity and well-defined project goals, pre-defined performance can help evaluate feasibility of the data science project early enough in the data-science life cycle. This early comparison helps the team refine hypothesis, discard the project if non-viable, change approaches.

8. How to efficiently scrape web data, or collect tons of tweets?

- Python example
- Requesting and fetching the webpage into the code: httplib2 module
- Parsing the content and getting the necessary info: BeautifulSoup from bs4 package
- Twitter API: the Python wrapper for performing API requests. It handles all the OAuth and API queries in a single Python interface
- MongoDB as the database
- PyMongo: the Python wrapper for interacting with the MongoDB database
- Cronjobs: a time based scheduler in order to run scripts at specific intervals; allows to bypass the “rate limit exceed” error

9. How to clean data?

1. First: detect anomalies and contradictions
   Common issues:

- Tidy data: (Hadley Wickam paper)
  column names are values, not names, e.g. <15-25, >26-45…
  multiple variables are stored in one column, e.g. m1534 (male of 15-34 years' old age)
  variables are stored in both rows and columns, e.g. tmax, tmin in the same column
  multiple types of observational units are stored in the same table. e.g, song dataset and rank dataset in the same table
  *a single observational unit is stored in multiple tables (can be combined)
- Data-Type constraints: values in a particular column must be of a particular type: integer, numeric, factor, boolean
- Range constraints: number or dates fall within a certain range. They have minimum/maximum permissible values
- Mandatory constraints: certain columns can't be empty
- Unique constraints: a field must be unique across a dataset: a same person must have a unique SS number
- Set-membership constraints: the values for a columns must come from a set of discrete values or codes: a gender must be female, male
- Regular expression patterns: for example, phone number may be required to have the pattern: (999)999-9999
- Misspellings
- Missing values
- Outliers
- Cross-field validation: certain conditions that utilize multiple fields must hold. For instance, in laboratory medicine: the sum of the different white blood cell must equal to zero (they are all percentages). In hospital database, a patient's date or discharge can't be earlier than the admission date

2. Clean the data using:

- Regular expressions: misspellings, regular expression patterns
- KNN-impute and other missing values imputing methods
- Coercing: data-type constraints
- Melting: tidy data issues
- Date/time parsing
- Removing observations

10. How frequently an algorithm must be updated?

You want to update an algorithm when:
- You want the model to evolve as data streams through infrastructure
- The underlying data source is changing
- Example: a retail store model that remains accurate as the business grows
- Dealing with non-stationarity

Some options:
- Incremental algorithms: the model is updated every time it sees a

new training example
Note: simple, you always have an up-to-date model but you can’t incorporate data to different degrees.
Sometimes mandatory: when data must be discarded once seen (privacy)
- Periodic re-training in “batch” mode: simply buffer the relevant data and update the model every-so-often
Note: more decisions and more complex implementations

How frequently?
- Is the sacrifice worth it?
- Data horizon: how quickly do you need the most recent training example to be part of your model?
- Data obsolescence: how long does it take before data is irrelevant to the model? Are some older instances
more relevant than the newer ones?
Economics: generally, newer instances are more relevant than older ones. However, data from the same month, quarter or year of the last year can be more relevant than the same periods of the current year.
In a recession period: data from previous recessions can be more relevant than newer data from different economic cycles.

11. What is POC (proof of concept)?

- A realization of a certain method to demonstrate its feasibility
- In engineering: a rough prototype of a new idea is often constructed as a proof of concept

12. Explain Tufte’s concept of “chart junk”

All visuals elements in charts and graphs that are not necessary to comprehend the information represented, or that distract the viewer from this information

Examples of unnecessary elements include:
- Unnecessary text
- Heavy or dark grid lines
- Ornamented chart axes
- Pictures
- Background
- Unnecessary dimensions
- Elements depicted out of scale to one another
- 3-D simulations in line or bar charts

13. How would you come up with a solution to identify plagiarism?

- Vector space model approach
- Represent documents (the suspect and original ones) as vectors of terms

- Terms: n-grams; n=1 to as much we can (detect passage plagiarism)
- Measure the similarity between both documents
- Similarity measure: cosine distance, Jaro-Winkler, Jaccard
- Declare plagiarism at a certain threshold

14. How to detect individual paid accounts shared by multiple users?

- Check geographical region: Friday morning a log in from Paris and Friday evening a log in from Tokyo
- Bandwidth consumption: if a user goes over some high limit
- Counter of live sessions: if they have 100 sessions per day (4 times per hour) that seems more than one person can do

15. Is it better to spend 5 days developing a 90% accurate solution, or 10 days for 100% accuracy? Depends on the context?

- “premature optimization is the root of all evils”
- At the beginning: quick-and-dirty model is better
- Optimization later

Other answer:
- Depends on the context
- Is error acceptable? Fraud detection, quality assurance

16. What is your definition of big data?

Big data is high volume, high velocity and/or high variety information assets that require new forms of processing
- Volume: big data doesn’t sample, just observes and tracks what happens
- Velocity: big data is often available in real-time
- Variety: big data comes from texts, images, audio, video…

Difference big data/business intelligence:
- Business intelligence uses descriptive statistics with data with high density information to measure things, detect trends etc.
- Big data uses inductive statistics (statistical inference) and concepts from non-linear system identification to infer laws (regression, classification, clustering) from large data sets with low density information to reveal relationships and dependencies or to perform prediction of outcomes or behaviors

17. Explain the difference between “long” and “wide” format data. Why would you use one or the other?

- Long: one column containing the values and another column listing the context of the value Fam_id year fam_inc
- Wide: each different variable in a separate column Fam_id fam_inc96 fam_inc97 fam_inc98

Long Vs Wide:
- Data manipulations are much easier when data is in the wide format: summarize, filter
- Program requirements

18. Do you know a few “rules of thumb” used in statistical or computer science? Or in business analytics?

Pareto rule:
- 80% of the effects come from 20% of the causes
- 80% of the sales come from 20% of the customers

Computer science: “simple and inexpensive beats complicated and expensive” - Rod Elder

Finance, rule of 72:
- Estimate the time needed for a money investment to double
- 100$ at a rate of 9%: 72/9=8 years

Rule of three (Economics):
- There are always three major competitors in a free market within one industry

19. Name a few famous API’s (for instance GoogleSearch)

Google API (Google Analytics, Picasa), Twitter API (interact with Twitter functions), GitHub API, LinkedIn API (users data)…

20. Give examples of bad and good visualizations

Bad visualization:
- Pie charts: difficult to make comparisons between items when area is used, especially when there are lots of items
- Color choice for classes: abundant use of red, orange and blue. Readers can think that the colors could mean good (blue) versus bad (orange and red) whereas these are just associated with a specific segment
- 3D charts: can distort perception and therefore skew data
- Using a solid line in a line chart: dashed and dotted lines can be distracting

Good visualization:
- Heat map with a single color: some colors stand out more than others, giving more weight to that data. A single color with varying shades show the intensity better

- Adding a trend line (regression line) to a scatter plot help the reader highlighting trends

visit: https://rpubs.com/JDAHAN/172473