THIS DOCUMENT CONTAINS QUESTIONS THAT REPRESENT THE SORT OF QUESTIONS THAT MIGHT APPEAR ON THE FINAL QUIZ FOR DATA MINING FOR BUSINESS ANALYTICS.

THESE ARE INTENDED TO REPRESENT THE FORMAT AND STYLE OF QUESTIONS, NOT NECESSARILY THE CONTENT.

Chapter-spanning Questions

1) Which data science method is most appropriate for the following business question? “Of all my accounts, which are the most likely to exhibit fraud, based on my experience with prior cases of accounts that have and have not been defrauded?”

   a) Classification tree induction

   b) Hierarchical clustering

   c) k-Means

   d) Linear regression

2) Which analytics technology would be most useful in answering the following business question? "If this customer responds to my offer, how much will she spend?"

   a) Classification tree induction

   b) Hierarchical clustering

   c) k-Means

   d) Linear regression

3) I’m proposing a foreclosure-classification system to a small bank with stakeholders of very varied backgrounds. For a pilot study demonstration on a small data set, which technique is most suitable?

a) Tree induction

b) k-Nearest Neighbor

c) Logistic Regression

4) Tree induction and clustering both can be used to segment customers. Contrast the process for using these two different types of modeling for customer segmentation. For what sort of problems would you use each? What would you have to do differently?

5) I want to rank credit applicants by their estimated likelihood of default. Which technique would be least helpful in assessing the quality of a ranking model mined from data?

a) holdout testing

b) calculate area under the ROC curve

c) calculate percent correctly classified instances

d) cross-validation

e) domain knowledge validation

6) I have a business application where I want to make fast classifications, and update my model quickly and immediately when a new training case comes along. What induction algorithm is best suited?

a) Naïve Bayes

b) Classification tree induction

c) k-Nearest Neighbor

d) Logistic Regression

Matching: Choose the *best* matching for each set; each letter should be used once (these might be hard)

_____cross-validation	a. Ranking
_____domain-knowledge validation	b. Comprehensibility
_____ROC curve	c. Generalization performance
_____overfitting avoidance	d. Complexity control

_____holdout evaluation	a. Sanity checking
_____domain-knowledge validation	b. Pruning
_____learning curve	c. Increasing data
_____overfitting	d. Cross-validation
_____increasing comprehensibility	e. Divergence between testing accuracies

______learning curves	a. increasing proportion targeted
______fitting curves for kNN	b. increasing tree size
______information gain	c. increasing training data
______cumulative response curves	d. increasing complexity
______increasing Mann-Whitney-Wilcoxon	e. increasing AUC

_____entropy	a. log odds
_____logistic	b. numeric target
_____information gain	c. how mixed up classes are
_____accuracy	d. higher on training data
_____regression	e. difference between parents and children
_____lift	f. better with model than without

Free-response Questions

Q) Blue Moon Consulting

After a few beers your CIO invited his buddy from Blue Moon consulting to propose a project using data mining to improve the targeting of the new service that you have been a principal in developing. The service has been quite successful so far, being marketed over the last 6 months via your ingenious, and very inexpensive, word-of-mouth campaign. You've already garnered a pretty large customer base without any targeting, and you've been seeing this success as your best stepping stone to bigger and better things in the firm. After some reflection, you've decided that your best course of action is to play a key role in ensuring the success of the data mining project as well. You agree with your CIO's statement in a meeting with Blue Moon, that very accurate targeting might cost-effectively expand your audience to consumers that word-of-mouth would not reach.

Identify the four most serious flaws in this abridged version of Blue Moon's proposal, and suggest how to ameliorate them. You can accept that Blue Moon has accurate information about the service.

"We will build a logistic regression (LR) model to predict service uptake for a consumer, based on the data on your existing customers, including their demographics and their usage of the service. We believe that logistic regression is the best choice of method because it is a tried and true statistical technique, and we can interpret the coefficients of the model to infer whether the attributes are statistically significant, and whether they make sense. If they are and they do, then we can have confidence that the model will be accurate in predicting service uptake. We will apply the model to our (Blue Moon's) large database of consumers, and select out those whom the LR model predicts to be the most likely to subscribe. It is a fixed-price, fixed-cost, fixed-term service, so this also will in effect rank them by expected profit as well."

Answer

1. Negative examples are not considered, as through WOM we only know who finally accepted (bought) the product. Also, due to WOM, the customers were probably not randomly sampled.
2. The choice of Logistic Regression is arbitrary.
3. Statistical significance of the coefficients is unrelated to the model's generalization performance.
4. The feature "usage of the service" will not be available for new customers. Therefore, we cannot use it in our predictive modeling.
5. It would seem that the current pilot is built on a totally separate dataset, not the one from BM. It's even unclear whether the information we currently rely on is available on BM's dataset.
6. It's possible that people would have taken the product anyway, and WOM has minimal effect ("friends" tend to have similar interests).

Q) A Political Campaign

Immediately after the successful 2012 Obama campaign, Campaign Manager and former White House Deputy Chief of Staff for Operations Jim Messina discussed their "Targeted Sharing" analytics program: "I think it's one of the most important things we did." Targeted Sharing (TS) addressed the problem that young voters were important for the Obama campaign, but were difficult to reach via the traditional channels employed by political campaigns (direct mail, phone, email). The TS solution involved targeting the friends of the roughly half-million "authorized" Facebook Obama supporters. The analytics involved prioritizing the friends for the "targeted" sharing of campaign content, as oversharing leads to Facebook deprioritizing the content in the friends' streams.

Your liberal leanings led you to volunteer, and now you have been selected to lead the design of an analytics method for prioritized targeting of the 70 million unique friends for a "getting out the vote" message. Choose one of the methods we have covered this semester. Explain your design by answering the following questions. The notion of "proxies" for unavailable data may be helpful. One or two sentences will suffice to answer each question.

1) What precisely is your problem formulation? What general category of data mining task does this correspond to?

2) What is your data representation? Is this a supervised or unsupervised formulation? Why? What features will you use? Describe a few elements of your feature vector precisely.

3) What method do you propose to use? Why?

4) How will you evaluate whether your model has captured any generalizable knowledge? Explain two different ways, and what metric you propose to employ.

5) If the evaluation shows that it has indeed captured generalizable knowledge, why will (help) to solve the prioritization problem? Explain precisely. Are there other important factors that should be included in the solution?

Answer

1) The key point here is that the target variable (who they voted for) may not be known ahead of time (prior to the elections; maybe even after the elections). So, either we take i) an unsupervised approach without a target variable or ii) use a proxy for our target variable (think of the pregnancy case).

2) Depending on the previous answer, the formulation is different. There are plenty of features to choose from. Make sure you are explicit and that the features can be (in the end) represented by a numerical or categorical value.

3) Again, this depends on how you are approaching the problem.

4) Think of the techniques we've seen to check for model generalization. For metrics, AUC is one. There are several ways to evaluate the model.

5) Tie your solution / proposal back to the business case. For example, do you care about all states? What if some states are blue states anyhow? Does it make sense to focus on these?

Q) You sell IT products and your lone data scientist proposes to use k-NN to build an IT wallet estimation predictor to predict the "wallet share" that companies spend with you. Wallet share is the percentage of a company's total IT budget that is spent with your company. You have information on the total IT budgets of a large set of companies, which will constitute your database of potential neighbors. The two of you have already gone over the distance measure and have chosen k. Now you want to estimate your wallet share for Acme Corp., one of your current customers for whom you do not know the IT budget.

a) Explain precisely how you will estimate your wallet share for Acme with this technique.
b) If you chose k=N, the total number of training examples, what would be the effect?

Q) Your data science team is presenting you the results of their latest analysis.

A) They present learning curves. Briefly compare and contrast learning curves and fitting curves, including what they show and why they are important.

B) They are asking for a budget to acquire more labeled training data. How can the learning curve help you to decide whether the request is justified?

C) They discuss results using cross-validation. What is cross-validation, and why would one use it?

D) In reporting their results, the team reports the area under the ROC Curve. What the heck is an ROC curve? When is it used and what does it show? Why should I care about the area under it?

Q) I would like to see whether my investment customers tend to cluster in understandable groups. What precisely is the difference between segmenting my customers using clustering and segmenting my customers using tree induction, and when would I use one rather than the other? What is the practical difference—i.e., in the data mining process where and what would the main differences be? What different clustering methods should I consider? Once I get clusters, it is important for me to see whether they make sense. Describe three ways to help understand the resultant clustering.

Answer

Clustering vs Decision Trees

Q) We would like to target some subset of the huge number of visitors to our main retail web page with a new special offer. Instead of the normal early May special offer of a discounted flower bouquet for Mom, we've decided to offer selected customers a 30% discount on any electric razor purchase from our stock. We need to decide which customers will get the normal special offer and which will get the new special offer.

(a) Show how computing expected value provides a framework for thinking about what models need to be built for this problem.

(b) Specify what models you would build.

(c) Do you expect to have the data necessary to build these models? If so, from where, if not, what do you propose to do about it?

We should compute the expected profit for a visitor from each of the offers, and offer each visitor V the offer that promises the highest expected value. Call the offers B(ouquet) and R(azor).

EP(V,B) = p(B|V)*v(B,V)
EP(V,R) = p(R|V)*v(R,V)

The costs of the offers can be built into the v() functions, since we don't incur costs unless the offers are taken. The p() functions are essentially different models that we use, one for B and another one for R, that estimate the probability of a visitor accepting the offer.

The v() functions would be estimated values from a customer responding to each offer. We may need models for these too, to estimate how much a Razor / Bouquet buyer would spend (what type of models would *these* be?)

To answer the question, we need to have data for **both** models on the normal May special offer. Therefore, we may need to gather data for one of the two cases (which one?). How would that be? What features would we be using to answer these questions? The features need not be the same for the two cases.

Q) Consider our telco churn example from class. We would like to target some subset of our customers with a brand new special offer prior to the expiration of their contracts, in order to minimize our losses due to churn. The special offer has a fixed cost of **C** irrespective of whether the customer accepts it. You might recall that we already did a similar expected-value based selection of customers in one of our homeworks. In real life however, you might want to design your solution more carefully and differently in some aspects.

(a) How would you approach this problem in reality from the point of view of expected value framework in order to select the subset of customers who should be targeted? Describe carefully your procedure.

(b) Assume you have a certain budget of $5,000,000 for your campaign. How would you decide which customers to target? How would you determine how many of these customers you will target?

(c) Specify what model or models you would suggest building, including specifying the target variable(s).

(d) Describe the types / groups of attributes you would consider; be as complete as you can while being brief.

(e) Do you expect to have the data necessary to build these models? If so, from where, if not, what do you propose to do about it? You can ignore any value of the customer if she leaves the company. You can assume that the value of a particular customer, if she were to stay, can be estimated directly using this particular customer's past value and knowledge of the business going forward and the incentive, so you don't need to use data mining to estimate this quantity.

Answer

a) Expected value framework can decompose the problem into subproblems that we can approach systematically and get values for, allowing us to address the high level problem through synthesis. We'd need 1) the expected value if we ***do not*** target the customer and 2) the expected value if we ***do target*** the customer. (Why do you think that is?) .
A similar setting appears in our textbook.

b) Following the answer to a), it appears that we need to build classification / class probability estimation problems. We would need different models, computing the following values:
   1. **p(stay | X, notI)** : the probability of the user to stay, given their characteristics and us ***not*** incentivizing them to stay
   2. **p(stay | X, I)** : the probability of the user to stay, given their characteristics and us giving them the incentive to stay (offer).

c) This is what we have been doing for defining target variables and so on. Any classification / class probability estimation model is acceptable.

d) We probably have a lot of data for b)-1) . We may need to collect data for b)-2).

Q) You have built a logistic regression model to help in your decision of which customers to target with a special offer prior to contract expiration. The model estimates the probability that a customer will leave within 90 days of contract expiration. You realize however that some customers are more valuable to you than others, and you can acquire data to quantify this. You have a budget of $10,000 to target customers with a special offer. Explain how you will use the results of your logistic regression model to target your customers.