1. **Many supervised learning algorithms "fit" numerical models to data. Explain in 2-4 sentences what this means and how it works.(1)**

Supervised algorithms such as classification, regression, and causal modelling, have a train set which contains some variables ‘x’ that we belief have some relationship with outcome known ‘y’, for which we want to use the ‘x’ to get a better idea of ‘y’. The objective functions among different methods are different, but the similar opinion is to get the best parameter to minimum the loss between the model predict value and the true one, such as we use SSE to compute the linear regression and mention hinge loss when it comes to linear SVM.

1. **Linear regression and logistic regression are two of the most common data mining algorithms. Explain in what sense they are similar and in what sense they are different with respect to the concept discussed in the previous question.(1)**

First, linear regression and logistic regression are both supervised learning algorithms, which means they both have the known ‘y’ as the outcome to fit the model.

They both use the same criteria to judge the usefulness, like R^2, and also should see the outlier/ high leverage point/ influential point, and should deal with multicollinearity problems.

The objective function of both two algorithms is SSE, while the outcome ‘y’ for linear regression is the original value compared to the log-odds of logistic regression.

Linear regression is dealing with continuous outcomes, is used in regression and prediction; while logistic regression is dealing with discrete outcome, used in classification (if family is Binomial distribution, it comes with 0/1; while Poisson distribution is non-negative integers).

Linear regression asks the variables to be normality, while logistic don’t strictly enquired.

The inner sense of both algorithms is nearly the same, only logistic regression add in a Sigmoid Function to map the y into a range.

1. **Explain briefly. What is overfitting? Why it is important? How do data mining algorithms control overfitting? Give an example. How can we tell if overfitting is happening? Be precise in your answers.(2)**

“Overfitting is the tendency of data mining procedures to tailor models to the training data, at the expense of generalization to previously unseen data points. ”(Textbook page 113) It is important as if we fit the training set so will, it will represent the unique patterns of the known data, which may not be useful to other unknown set, or to said it is allowed to pick up harmful spurious correlations. So, in the data separation part, we can also separate the known dataset into training data and holdout data, so that after we train the data and get the model, we can use holdout set to choose the most appropriable hyper-parameter in the model to balance the problem.

One of the example is to form the holdout evaluation to cross-validation, split the data into k different folds, and the iterate the training and testing k times to avoid the single lucky pick and so as to get the average performance and choose the best parameter.

We can see that when increasing the complexity, while the training set’s error goes monotony down, the holdout data’s performance will go up and the down. Then when the test error get higher again, it means we may have overfitting problem.

Diagram

Description automatically generated

1. **Naïve Bayes is a popular data mining algorithm, but it assumes conditional independence. In contrast, it is not necessary to make this assumption when estimating a decision tree. If the conditional independence assumption is NOT met, is it possible for a Naïve Bayes model to make better predictions than a decision tree that was estimated using the same data? Justify your answer.(1)**

My answer is possible. As George E.P. Box said, “All models are wrong… Some are useful”. Even the conditional independence assumption is not met, the Naïve Bayes still give a sense of the data. Except the conditional independence assumption, it will use the probability to generate a convincing result. One of the occasion is that, when facing an extremely high range of variables (like shopping list, or words), The decision tree to choose some variables’ specific range to split will be meaningless, as it will easily stuck in a local minimum point, have a huge overfitting problem and can’t have a good application to estimate other given data. Even apply to the random forest, sill will overfitting when the samples are large-sized with great noise.

1. **Plumbing Co. has sold plumbing supplies for the past 20 years. The owner, Joe, has decided that next year he will diversify his business by adding gardening tools to his products. Joe has been successful using customer data to build predictive models and decide who to target with special plumbing offers. He believes he could use data mining to identify a subset of customers who would be good prospects for gardening tools. Is Joe prepared to solve this as a supervised learning problem? If yes, what would you suggest as the target variable? If not, explain why not and suggest a possible course of action for Joe to achieve his business objective.(2)**

As this gardening tools are not existed in the previous year, it is a new experiment and thus I think use previous data may be unreasonable. Plumbing supplies contains many parts, and the potential customer of gardening tools may not intercept well with other supplies. But this doesn’t mean he have no idea or target to all the person.

There are several actions he can do. First, he can find other proxy data, such as the competitors who is already have this tool and see the distribution and patterns of this tool’s target customers. Second, to do the research and focus group interview. Collect the information of the background, ask several aspects, and discuss for the advice. Thirdly, randomly pick a sample call them to get a brief concept of the customer construction or do the A/B testing to see the result.

But also, if this is an urgent condition, or simply do the assumption that the tool is similar to others, and just collect the existing data to do the supervised learning would be fine. But still recommend using others mentioned above to get a more comprehensive idea.

1. **You have been hired as a Gotham Insurance (GI) risk analyst to lower the cost of life insurance products. GI has noted that several of its life insurance customers have died due to cancer recurrence after suffering from breast cancer. Therefore, GI is assessing the possibility of giving a free medical follow-up to life insurance customers who suffered from breast cancer but were cured. Hopefully, this will reduce the likelihood of death due to cancer recurrence. You already have a model to predict the probability of cancer recurrence for a given person. Now, you need to decide to whom GI should give free medical follow-ups.**

**Here are a few additional details about the context:**

* **When a customer with life insurance dies, GI’s life policy pays $200,000.**
* **Early detection of cancer recurrence decreases the probability of dying by 10%.**
* **A follow-up after cancer recovery costs $5,000.**

**Based on this information, we should offer a free medical follow-up to individuals who have a probability of cancer recurrence greater than what? Use a decimal number with two decimal spaces for your answer. For example, if you think the answer is 10.7%, your input should be 0.11.(1)**

0.03

1. **You are an analyst that works for the Customer Service Center (CSC) of Nosara, a company that sells clothes online. CSC employees are responsible for responding emails about returns and exchanges, fashion trends, and complaints. Every day they receive hundreds of emails to which they would like to give different priorities, especially complaints (which fortunately are few). They want you to use data mining to classify these emails.**
2. **Suppose you were tasked to build a proof of concept. Describe in detail (at least a paragraph) how you would proceed and why.**
3. **A more complex version of your preliminary solution could include benefits and costs. Describe (in at least 2-3 sentences) how you would structure the benefits and costs for this problem.(2)**
4. If the worker/ internship position is enough, first spend some time in message processing: ask a group of people to go through a large enough sample size of data, detect the massage’s meaning and then separate into the groups. After the training data have some kinds of patterns, using bag-of-words and fit with the supervised learning algorithms such as logistic regression and find the probability of the complained message (combined all others to be ‘non-complain’, and group others when fitting another one). But if there’s no such time and labours, maybe try using the existing packages such as Python package ‘TextBlob’ to detect the simple text processing, for example, find the correlated words such as ‘service’ , ‘customer’ for complaint etc. and see what’s going on.
5. Take ‘complaint’ as an example. If you target that complained customer, and solve his/her problem successfully, he/she will continue buying the goods, if we treat to be time-independent (take Poisson), then the benefit equals to the average money people spend on this shop. And the cost will split into two part: first is the general cost, like the money and time spend by constructing the initial database, the money to maintain and improvement of the algorithm; another part is the time and money to deal with their individual problems, such as hire people to negotiate or give customer discount etc.
6. **You are responsible for detecting fraud in an online store; fraud occurs when a scammer logs into a customer's account and purchases digital products. This is expensive for the store because it must pay its suppliers for the cost of the stolen goods. You can model quite well (1) the probability that an account has been compromised and (2) how much you would lose with a particular account if it were actually compromised. On the other hand, management is concerned with upsetting and alarming customers, so you must be careful when deciding whether to take corrective action (e.g., force password change, verify transactions). What would you do to decide whether to take corrective actions? Be specific in your answer.(2)**

As we know the probability that an account has been compromised, so from the account responsible opinion, you should detect the group of people which have very simple password, easy-detected transaction method and so on even you will not loss much money while losing it (but if the cost is high, the threshold can also set high), and if know the customer’s tolerance of the reminders, can change flexible.

And then, it is a benefit-cost problem. If we can know the individual’s behaviour (like the probability difference) that whether take corrective actions or not, then make a comparison of the cost and profit and choose the smaller one.

the profit = ‘the difference probability’ multiple by ‘the lost if it were actually compromised’

the cost = of ‘the reminder setting’ plus ‘the probability of leave forever due to the frequent reminder’

1. **You are working for GloboBank, one of the world's largest financial institutions. They want to build a new system to monitor salespeople’s electronic communications with customers. The goal is to reduce bad behavior among the company's salespeople, such as overpromising, understating risks, and so on. The company is unhappy with their current surveillance system because it creates tons of false alarms, which wastes the time of the analysts. They also know that it misses a lot of important cases. Below is a proposal they have received from a vendor, for a better system, that should address these issues. Specifically, it will monitor each outgoing email from a salesperson and flag those that are suspicious. The flagged-suspicious emails would then be examined by an analyst, who would decide which ones ought to be escalated for further investigation. Assess the proposal and provide constructive criticism: identify the three most important potential flaws and suggest ways to fix each of them.**

**Proposal from Yellow Fin Consulting**

**We will use machine learning techniques to build a system to classify emails as “suspicious” or not. Those classified as suspicious will be “flagged”; our surveillance system will rank emails and provide the most suspicious ones to the analysts. The system will maximize the lift at the top of the ranking and minimize the number of missed cases (false negatives).**

**The flagging model will take as input a feature-vector representation of the email, where each word is a binary feature that represents whether the word is present in the email or not; more sophisticated representations will be added later. We will leverage the existing system to provide training labels. Specifically, if the existing system flags the emails as being suspicious, we will give them a label of yes. Otherwise, we will give a label of no. As we archive all salesperson emails specifically for compliance purposes, this will allow us almost unlimited training data.**

**We will evaluate the system based on its generalization accuracy and the area under the ROC curve (AUC) on holdout data. The system should be able to achieve accuracies greater than 90% as well as high AUCs. We also will show the flagged emails to compliance experts for domain knowledge validation.(3)**

1. As this proposal use all emails as training set and generate the model, one of the main concern is the overfitting problem. Even the data is from all salesperson, maybe exist some bias (like if the holdout data separate by time, maybe there may exist some time-dependent problem under detected), or the model fit the data so well such that exist some harmful spurious correlations. So my advice is do the cross-validation and other methods to find the balanced point between error and complexity.
2. Like the proposal mentioned, they used each word as binary feature and put them into model, but in some occasion, word order matters more. For example. “Wait, not good” and “No wait, good” contains exactly same words but reflect different meanings. So to solve the problem, apply more advanced method such as such as the n-gram model to combine the sequence into the model would be appropriable.
3. This proposal use AUC as only result of the model evaluation, which is much less efficient. Even the AUC is more than 0.9, but what about only a very few of the salesperson are suspicious? Thus the model gives every a “NO” tag, which have a high accuracy but the model is completely meaningless. In order to deal with this problem, maybe add in more score such as ‘the percentage of predict no-suspicious while this person is suspicious’ to get a better idea. Or try different methods, such as Naïve Bayes, use different opinion to solve the problem.