Applied Machine Learning Mid-term

Part I: Applied ML Basics

Solve any 2 out of diabetic retinopathy detection, ads recommendation and relevance ranking.

You are given a description of a situation. You are supposed to come up with:

- Problem Framing
- Why use ML
- Solution Design

You may make any reasonable assumptions (if needed) and explicitly state your assumptions and reasoning. You may use the templates provided as a guideline.

1. diabetic retinopathy detection

   Consider a hypothetical situation in which you are asked to provide a decision support for a hospital by building a model to detect severe diabetic retinopathy condition from the retina images of the patients. These patients will be further referred to a specialist doctor in the hospital for a detailed evaluation and a potential surgery. On an average the hospital receives 10000 patients in a year. Although the hospital has many non-specialist staff, they only have 2 specialist surgeons. The detailed evaluation and surgery is time consuming. One surgeon, on an average, can only evaluate around 50 patients a year. The prevalence rate of severe diabetic retinopathy among the patients that hospital receives is around 10%. Currently the hospital has non-specialist staff who do the preliminary examination of patients and refer them to the specialist. This process is time consuming. Also the non-specialist staff frequently miss the subtle clues that indicate the presence of the condition, hence losing a potential customer. The hospital wants to improve their revenue by detecting and treating as many patients with severe diabetic retinopathy as possible. At the same time the hospital can not afford more than 20% false detections as it can affect their reputation.

2. ads recommendation

   Consider a hypothetical situation in which you are asked to improve the recommendation engine of an online advertisement platform. The advertisers bid to pay a certain amount for a click on their ad. The platform gets paid the bid amount only if the user clicks on the advertisement. The non-clicks do not generate any revenue. Currently the platform is using a simple strategy of showing top 5 ads with the highest bids. The platform suspects that this strategy is not optimal . Also it is creating a bad user experience by showing them irrelevant ads. They want to improve the revenue by building a model that can predict the probability of a click based on historic data and use this probability together with the bid amount to rank their ads and show top 5 ads. They want, on an average, at least 2 out of 5 ads to be relevant in order to maintain a decent user experience.

3. relevance ranking

Consider a hypothetical situation in which you are asked to improve the performance of a search engine which gives relevant results to a search query. Currently the search engine is using a keyword based matching. The main concern is that the mobile users are finding irrelevant results displayed on first page and hence have to scroll down multiple pages to find a relevant result. This causes a bad user experience. During evaluation of user experience, several thousand examples user and search result are manually labelled as relevant or not relevant. So the goal is to build another better model on top of the current model. The two layer model will first use the current model to get the top 30 results. Then it will predict the relevance of these 30 results using the second, much better model and rank them so that the top 5 results do not contain many irrelevant results.

Part II: MLOps

1. data
   a. profiling
      What statistics would you use to profile the following:
      - a numerical column : (e.g. salary)
      - a categorical column: (e.g. gender)
   b. cleaning
      What strategies would you use to clean the following:
      - a numerical column: (e.g. age)
      - a categorical column: (e.g. city name)
   c. drift detection
      What strategies would you use to detect data drift for the following:
      - a numerical column: (e.g. salary)
      - a categorical column: (e.g. gender)
2. model
   a. overfit check
      - How would you detect if a model is overfitting?
      - How would you fix it?
   b. underfit check
      - How would you detect if a model is underfitting?
      - How would you fix it?
   c. threshold adjustment
      - Suppose we want to increase the recall of the spam classification model. Would you increase the threshold or decrease it?
      - Why?
3. code
   a. unit test: Write a unit test for a function to compute the mean value of an array. Specify 2 inputs and expected outputs covering typical and edge cases.

b. automation: What tool can you use to automatically run the unit tests when someone tries to commit the code to the master branch in git?

c. deployment: What tool can you use for smooth and reproducible deployment or migration from one machine to the other?

Part III: Statistical Learning Basics

1. methods
   a. classification
   Give an example of a problem where you would use classification.
   b. regression
   Give an example of a problem where you would use regression.
   c. clustering
   Give an example of a problem where you would use clustering.
2. quality of models and limitations
   a. quality of fit: How would you measure quality of fit of a regression model?
   b. bias-variance trade-off: What is bias variance trade-off?
   If the bias is high and variance is low does it mean the model is overfitting or underfitting?
   If bias is low but variance is high does it mean the model is overfitting or underfitting?
   c. limitations: irreducible error and bayes error rate
   What is irreducible error in regression?
   What is bayes error rate in classification?
3. statistical inferences
   a. confirming relation: How do you determine if there is a linear relation between a predictor variable and target?
   b. variable importance: Suppose we have trained a linear regression model:
   sales = 2.9 + 0.05 x TV + 0.19 x Radio - 0.01 x Newspaper
   which advertising channel has the most impact on the sales? Why?
   c. extrapolation error: Suppose you have the data for the average heights of 4 to 6 year old boys

age	4	4.5	5	5.5	6
height(cm)	109.7 102.9	109.7 106.2	109.9 109.7	113.4 109.7	116.1 109.7

−6.8 −3.5 0.2 +3.7 +6.4

1
3.6
6.4
6.8
16.8

~~+~~6.8+~~+~~1.75 +
1.85+ 6.4
3.6

$$\frac{1^2+0.5^2+0^2+0.5^2+1^2}{\boxed{2.5}}$$

Suppose you want to use linear regression model
height = alpha x age + beta
what value of alpha and beta would you choose?

What will be the predicted height of a 50 year old person based on your model? What went wrong?

2 2
102.9
106.2
109.9
113.4
116.1
548.5
5

4 1
6.72
× 6
40.32

~~416.7~~ 5 4
−102.9
13.2

$\frac{13.2}{2}$

91.1 + ~~6~~
40.32
109.7
131.42

3 1
6.72
× 5
18.60

6.6
× 0.5

$\frac{16.8}{2.5}$

$\hat{\beta_0}$ = ~~$\bar{y}$~~ 109.7 − (18.6)

102.9
+ 3.3
106.2
3.3
109.5

6.72
25) 168
~~(150)~~

1
10¹9.9
+ 3.3
113.2

0 9 19
~~1~~109.7 − 18.6
91.1

~~113.4~~
+ 3.3
116.7

~~25) 168~~

~~91.1~~
~~+ 18.6~~
~~109.7~~

5

109.7 − 33.6
76.1

$\hat{\beta_1}$ =

76.1
+40.32
116.42

6.72

91.1

1 1
76.1
336
412.1

$$\hat{\beta}_1 = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2\ ~~\Sigma(y_i-\bar{y})~~^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

3 1
6.72
× 5
33.60

16800

Templates

Problem Framing

	qualitative answer	quantitative answer	question
Current State			what is the current situation that we want to address and why?
Objectives			what is that we want to achieve and why?
Benefit/ Cost Tradeoff and Prioratization			what are the cost of errors/benefits of correct predictions and why?
Constraints			what are the constraints/acceptable risks and why?
Desired State			what is the desired outcome (benefits/costs) that we want to see and why?

Why ML

	qualitative answer	quantitative answer	question
best non-ML alternative hypothesis			what are the non-ML alternatives and why are they problematic? (pains/missed gains)?
ML value proposition hypothesis			what are the advantages (pain relievers/gain creators) of ML solution and why?
ML feasibility hypothesis			what data and model are good candidates and why?

Solution Design

	choices	metrics	experiments
data			
model			
action			
reward			