Dhirubhai Ambani Institute of
Information and Communication Technology

Surprise Quiz I

IT496: Introduction to Data Mining

Date: 22nd September 2023 | Timings: 03:40 PM – 04:00 PM

READ THE TEXT BELOW CAREFULLY

- *Note that many questions have multiple correct answers.*
- For each question, write what you think is (are) the correct option(s) and support your answer with **brief** and **justified** explanations in 1-2 sentences in your answer sheets.
- For each question, you will be given **3** points for the correct answer (only if you select all the correct options with the correct reason) and **-1** point *otherwise*.

10 questions | **30** total points | **20** minutes | **+3** if correct **-1** otherwise.

Suppose you are using Polynomial Regression. You plot the learning curves and notice a large gap between the training and validation errors.

1. What is happening?

   (a) Generalization Error (b) Sample Error
   (c) Underfitting (d) Overfitting

   This is likely because your model is overfitting the training set.

2. How to solve this?

   (a) Increase the polynomial degree and regularize the model using L_1 penalty
   (b) Decrease the polynomial degree and regularize the model using L_2 penalty
   (c) Increase the polynomial degree and the size of the training set
   (d) Decrease the polynomial degree and increase the size of the training set

   Simplify the model and gather more training examples.

Suppose you are using Ridge Regression and notice that the training and validation errors are almost equal and fairly high.

3. What is happening?

   (a) High bias
   (b) Low variance
   (c) High bias, low variance
   (d) Low bias, high variance

   If both the training and the validation error are almost equal and fairly high, the model is likely underfitting the training set, which means it has a high bias.

4. How to solve this?

   (a) Use Lasso Regression with a low value of λ
   (b) Use Ridge Regression with a low value of λ
   (c) Use Lasso Regression with a high value of λ
   (d) Use Ridge Regression with a high value of λ

   Stick to your existing model but remove constraints to increase its complexity.

---

5. Suppose you want to classify pictures as *outdoor/ indoor* and *daytime/ nighttime*. How would you solve this problem?

   (a) Implement two Logistic Regression classifiers
   (b) Implement one Softmax Regression classifier
   (c) Implement a one-versus-rest Logistic Regression classifier
   (d) Implement a one-versus-one Logistic Regression classifier

   Since these are not exclusive classes (i.e., all four combinations are possible), you should train two Logistic Regression classifiers.

---

6. Select the ***True*** Statements about the gradient descent (GD) algorithm from the following.

   (a) Stochastic GD has the fastest training iteration; however, it does not converge to the global optimum

   (b) Batch GD generally needs higher training time; however, it converges to the global optimum

   (c) Mini-batch GD bounces around the optimum unless we gradually reduce the learning rate

   (d) Stochastic GD may get stuck in a local minimum while training a logistic regression model

(e) It is a good idea to stop Mini-batch GD immediately when the validation error goes up

SGD has the fastest training iteration since it considers only one training instance at a time, so it is generally the first to reach the vicinity of the global optimum (or Mini-batch GD with a very small mini-batch size). However, only Batch GD will actually converge, given enough training time. As mentioned, stochastic GD and Mini-batch GD will bounce around the optimum unless you gradually reduce the learning rate.

---

7. You are given a noise-free dataset **D = {(y,x)}** consisting of `{(true,1),(true,2),(true,4)}`. You claim to have an unbiased binary classifier that predicts label probabilities. What probability in **[0, 1]** (*closed interval of range*) would you predict for the following cases?

   (a) P(y=true | x=2)
   (b) P(y=true | x=3)

   (a). P(y=true | x=2) = 1; noise-free
   (b). P(y=true | x=3) = 0.5; unbiased, completely uncertain

---

For the following tasks, which boolean or ranking metric would you evaluate on test data to compare the performance of systems, and why? Assume there is labeled test data.

8. Classification of GSRTC bus wheel images as faulty (e.g., cracked) or safe.

   Recall / Recall@k / F-measure
   This is a safety issue; therefore, we need all faulty images.

9. Classification of whether patients discharged from a hospital will be readmitted in the next month. The top 50 most probable readmission cases are selected for follow-up appointments; no action is taken for the remaining patients.

   Precision@k | recall@k, where k = 50
   The measure cannot be boolean.

10. Patent retrieval in an information retrieval setting where the user has to find all patents that could invalidate the novelty of a new patent application.

    Recall / Recall@k / Precision@k (if time is limited)

***** End of the Paper *****