# Machine Learning Assessment 12

Question 1

In his social network company, data scientist John Doe needs to build a classifier to detect hated speech. He collected data from users' posts and labeled them from community flags. The data has 3000 columns, he ran PCA to reduce to 2 dimensions and visualize 2 classes. To improve model performance what metrics should he optimize for?
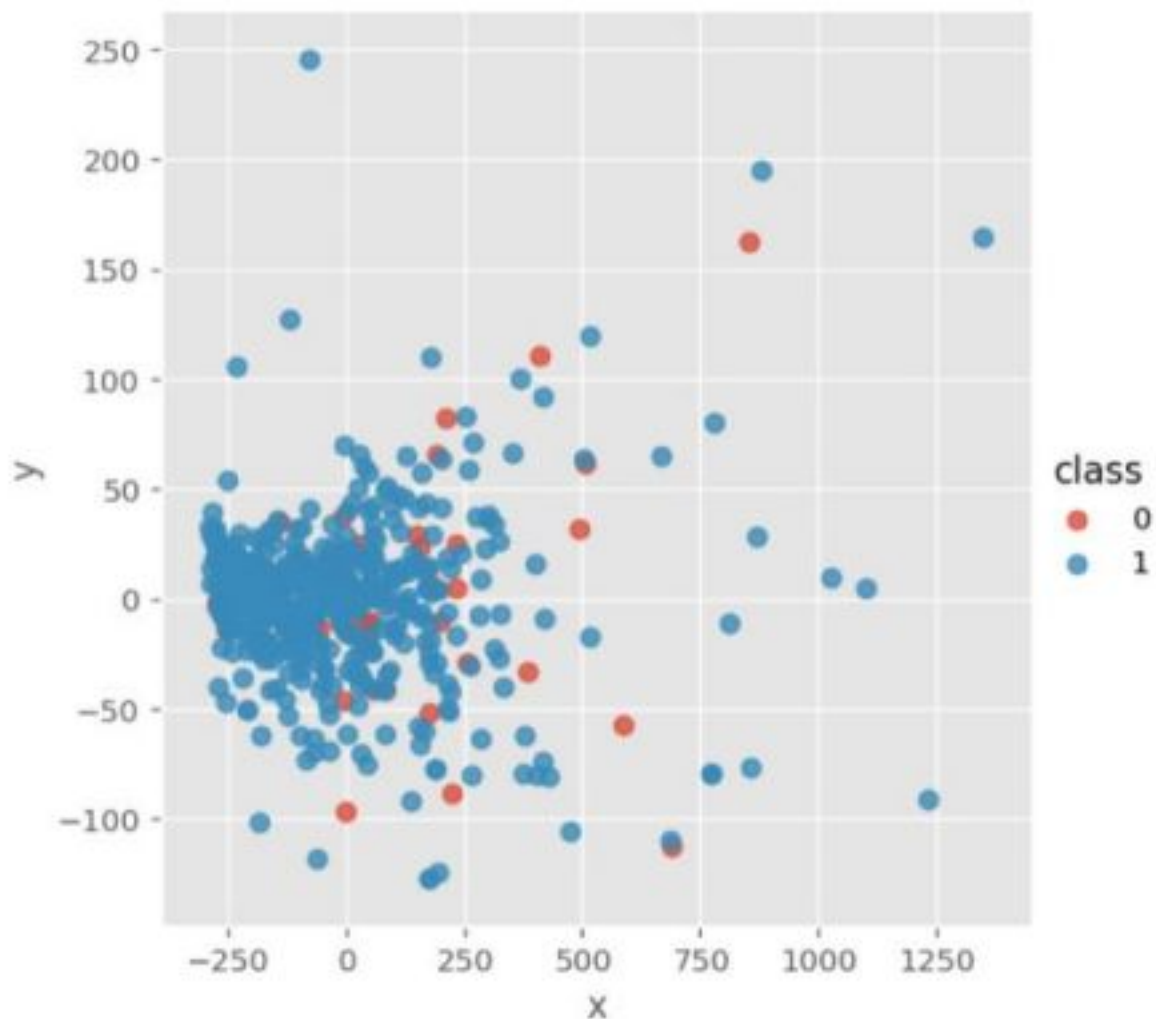
Answer

It's important to realize there is imbalance class distribution, class 1 is a lot more than class 0. Therefore we need to pay attention to the metrics you're optimizing for. AIC/BIC tend to penalize complex models more heavily, giving preference to simpler models in selection. See Elements of Statistical Learning page 233. If you got this question wrong, pay attention to the model assessment and selection; How to handle imbalance class (search smote imbalance learning).

Matthews correlation coefficient is more robust to imbalance class however it's not easy if you want interpretation. Different companies usually favor a different approach to data science as such when being asked you can mention any metrics you knew about and their pros/cons. The wrong way to answer is to insist your preference is the right way.

Question 2

John wants to choose some features for his model. One common idea is to use features important from random forest. John then run this function: get_feature_importance(full_data_set). Do you agree with this method?

Answer

If you failed this question, you can revisit if you make the same mistakes in other modelling steps. To make models more generalizable and robust top unseen data, we want to preprocess data on training data exclusively. The right answer is to perform feature importance on training data only, avoid the risk of information leakage. Information leakage is very common and

easy to make in practice.

Question 3

Once John finishes building the classifier, the business team is happy with reasonable performance on the train set but they see this report for the test set and they are concerned because a lot of non-hatred posts are detected as hatred. What would you recommend John to verify.

		Predicted	
		normal	hatred
Actual	normal	600	400
	hatred	10	100

Answer

This is a confusion matrix. Remember in question 1, we're dealing with imbalance class. In this confusion matrix, the distribution is 1000/110, in other words normal class is 10 times more than hatred class. However the model predictions seemed to result in roughly 610/500 or about 50/50. Therefore one needs to wonder if it should not be a coincidence and should check the distribution in the training data set. This question wants to test observability and actual hands on experience.

Question 4

Random forest able to reduce variance because (1) Training on different samples of data (2) Use random subset of features (3) All of the above.

Answer

The right answer is (3) All of the above. This is purely a knowledge checking question. A different way to ask this question is how random forests reduce variance. Almost all companies ask questions about random forests.

Question 5

If you flip a coin 50 times and try to get heads, what is the variance of the binomial distribution?

Answer

12.5 you can use the formula of binomial distribution to get the answer.

Question 6

Which of the following hyperparameter(s), when increased may cause Random forest to overfit the data? (1) Number of Trees (2) Depth of Tree (3) Learning Rate

Answer

The correct answer is depth of tree. You can read more about random forest from the Elements of Statistical Learning. You should also know that there is no learning rate in Random Forest.

Question 7

For which of the following hyperparameters, higher value is better for Decision tree algorithm? (1) Number of samples used for split (2) Depth of tree (3) Samples for leaf (4) Can't say.

Answer

This is a trick question, the right answer is can't say because it's depended on what you're optimizing for your use case.

Question 8

After fixing all the issues with his classifier, John persisted in his scikit learn model pipeline as a pickle file and deployed it in the server. What could possibly go wrong? (1) Pickle might have a compatible issue with different scikit learn version (2) It won't work because pickle file doesn't persist preprocess steps.
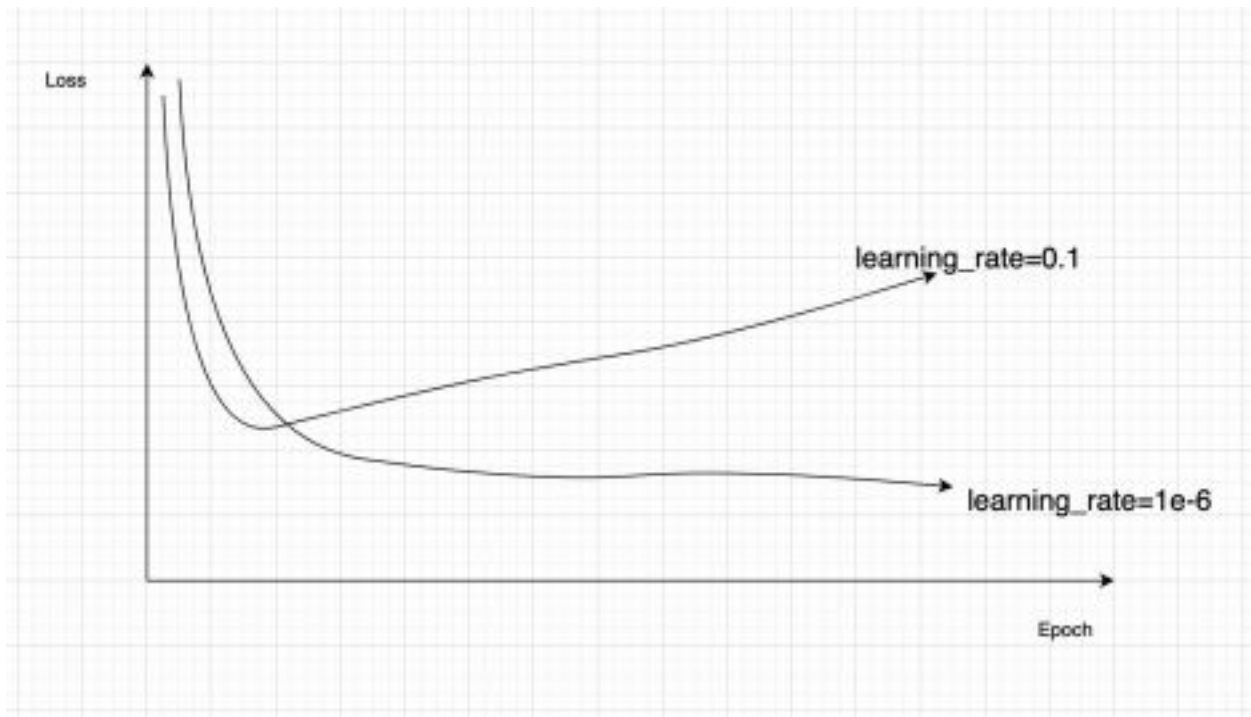
Answer

The right answer is (1). When you pickle a pipeline in scikit learn you can use it as-is in serving new samples. It can persist the whole pipeline.
Scikit
Learn is very. popular machine learning toolkit even in industry. If you got this question wrong you missed a chance to show that you have built machine learning in a production environment.

Question 9

John wants to improve his model and he uses deep learning. When he plots training loss for different learning rates he sees this chart. What learning_rate should he try to get better performance?

Answer

This is a classic example to check if you can distinguish high learning rate vs low learning rate. As you can see learning rate 0.1 cause the loss drop sharply and increases very quickly as you trained more epochs. It's indicated high learning rate. On the other hand, learning rate 1e-6 cause loss reduce very slowly. It's indicated low learning rate. The better learning rate can be achieve by value in between. Andrew Ng have the best explanation on youtube (Machine Learning on coursera).

Question 10

John thinks deep learning can get better performance so he built 3 layers of

neural networks. What else would you recommend him to start? (1) Use relu activation for all layers, use mean square error as loss function, if output > 0.5 assign it as class 1. (2) The last layer he uses is sigmoid, the output then will be used as probability of class 1. (3) The last layer he uses tan, the output then will be used as probability of class 1.

Answer

The right answer is (2). Sigmoid is very useful for binary classification because its value is bounded within [0,1]. If you want to review you can see any Keras tutorial for binary classification. In practice choosing the right activation function, loss function are crucial. The classic textbook DeepLearing is the best resource for deep learning fundamentals.

Reference:
   - Original quiz