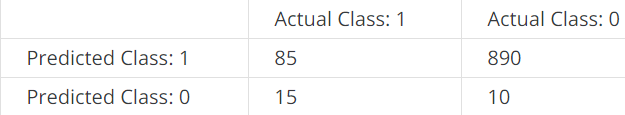
**Machine Learning System Design**

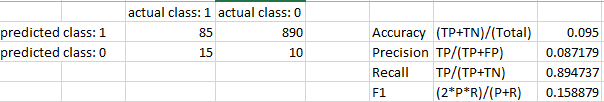
* You’re working on a spam classification system using regularized logistic regression. "Spam" is a positive class (y = 1) and "not spam" is the negative class (y = 0). You have trained a classifier + there are m = 1000 examples in the cross-validation set. The chart of predicted class vs. actual class is:

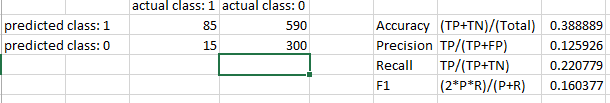


For reference:

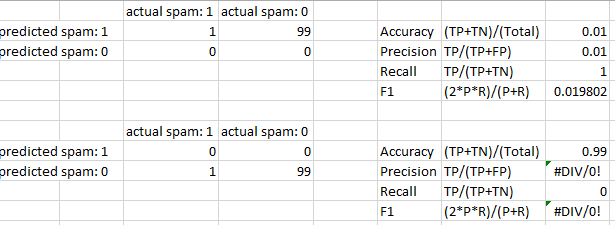
* Accuracy = (true positives + true negatives) / (total examples)
* Precision = (true positives) / (true positives + false positives)
* Recall = (true positives) / (true positives + false negatives)
* F1 score = (2 \* precision \* recall) / (precision + recall)

What is the classifier's F1 score (as a value from 0 to 1)?

* **0.158879**
* 
* Suppose a massive dataset is available for training a learning algorithm. Training on a lot of data is likely to give good performance when two of the following conditions hold true.
* **The features x contain sufficient information to predict y accurately. (For example, one way to verify this is if a human expert on the domain can confidently predict y when given only x).**
* **We train a learning algorithm with a large number of parameters (that is able to learn/represent fairly complex functions).**
* Suppose you have trained a logistic regression classifier which is outputting hθ(x). Currently, you predict 1 if hθ(x) ≥ threshold + predict 0 if hθ(x) < threshold, where currently the threshold is set to 0.5. Suppose you increase the threshold to 0.9. Which of the following are true? Check all that apply.
* **The classifier is likely to now have higher precision + lower recall 🡪** more FN’s, less FP’s = better TP Rate
* Suppose you decrease the threshold to 0.1. Which of the following are true?
* **Classifier is likely to now have higher recall 🡪** more FP’s, less FN’s, less TP’s = better TN Rate



* Suppose you are working on a spam classifier, where spam emails are positive examples (y=1) and non-spam emails are negative examples (y=0). You have a training set of emails in which 99% of the emails are non-spam and the other 1% is spam. Which of the following statements are true? apply.
* **If you always predict non-spam (output y = 0), classifier will have an accuracy of 99%.**
* **If you always predict spam (output y = 1), classifier will have a recall of 100% + precision of 1%.**
* **If you always predict non-spam (output y=0), your classifier will have a recall of 0%.**



* Which of the following statements are true? Check all that apply.
* ***It is a NOT good idea to spend a lot of time collecting a large amount of data before building your first version of a learning algorithm.***
* ***If your model is underfitting the training set, then obtaining more data is NOT likely to help.***
* ***After training a logistic regression classifier, you don’t HAVE to use 0.5 as your threshold for predicting whether an example is positive or negative.***
* **The "error analysis" process of manually examining the examples an algorithm got wrong can help suggest good steps to take (e.g., developing new features) to improve performance.**
* **Using a very large training set makes it unlikely for model to overfit the training data**.
* **On skewed datasets (e.g., more positive than negative examples), accuracy is not a good measure of performance + you should instead use *F*1 score based on the precision + recall.**