Damien martin Data science

When interviewing for data science positions, you are likely to be asked what sort of metric you would use in a classification problem. If a product manager is asking you this question, then the answer is almost certainly more nuenced than "precision" or "recall". The problem (as discussed here) is that "precision" and "recall" are more academic in that they prioritize the performance of the *model*, rather than the actual business problem which may require a more nuanced metric.

If you are being interviewed by a data scientist or a machine learning engineer, however, they may want you to use the terms "precision" and "recall". Part of this is because they are easy metrics to use in cross-validation (both recall and precision are built-in), part of it is because they are used frequently in articles and blogs. A small part is probably "if I had to learn this, you should know it too"!

Here we will assume that you are being asked a question by a data scientist or ML engineer, and give some tips on how to structure your response. Before starting, here are the definitions of precision and recall as a quick reminder:

Term	Description
Precision	Fraction of actually positive cases among those predicted to be posit
Recall	Fraction of actually positive cases found from all positive cases
True positive rate (TPR)	Same as recall; fraction of positive cases found

False positive rate (FPR)	Fraction of actually negative cases identified as positive

A previous article, *Bad Names In Classification Problems*, also defines precision and recall (and gives some alternatives to the names TP/FP/TN/FN). The TPR and FPR are the vertical and horizontal axes on the ROC curve, respectively.

The short version

Recall is more important where Overlooked Cases (False Negatives) are more costly than False Alarms (False Positive). The focus in these problems is finding the positive cases.

Precision is more important where False Alarms (False Positives) are more costly than Overlooked Cases (False Negatives). The focus in these problems is in weeding out the negative cases.

Sample interview question

An interview question might be

> *We are trying to build a classifier to identify whether a particular insurance claim is fraudulent. What do recall and precision mean in this context, and which is a better metric for our problem? Why?*

Don't answer this question by repeating the definition of precision and recall back to the interviewer. Telling her that "precision is the fraction of actually positive cases from those our classifier labeled as positive" is going to ding you on communication. Instead, identify what the positive case is in this example, and explain what the terms would mean *in this context*.

Since we are also asked which we should prioritize, we should also explain what the consequences of having a low precision or recall score are *in this context*. We focus on the low case as it is always good to have precision and recall as high as possible; by talking about what the problems with a low precision or recall are allow us to discuss the tradeoffs in terms relevant to the business.

Using this example, we are trying to detect fradulent claims. This makes the "positive case" (the thing we are trying to detect) the fraudulent claims. Giving our definitions in terms of fraud vs not-fraud (instead of positive and negative) gives us the following:

- **Precision** is the fraction of cases that are fraudulent out of those that our classifier labelled as positive. The problem with a low precision is that our investigators will spend a lot of time investigating claims that are actually legitimate.
- **Recall** is the fraction of fraudulent cases our classifier finds. The problem with a low recall is that we would be paying out on a lot of undetected fraudulent claims.

This would give you a basis for discussing how to trade recall vs precision. In this example, paying a claim you didn't have to is more expensive than paying someone to investigate the claim, so recall would be more important. The easiest way to maximize recall is to flag *every* claim, which clearly makes the classifier useless, so even though recall is more important, phrasing it this way can help determine how you would evaluate a trade-off between precision and recall.

Sample problems

To help get some experience with this, try answering the following questions by rephrasing the meaning of precision and recall in the context of each problem. When doing this, also practice evaluating what the downside of having too low a precision or recall is for that problem. Answers are available (but start collapsed).

Problem 1: Disease detector

1. What is the positive class?
2. What would a recall of 80% mean?
3. What would a precision of 75% mean?
4. If the recall is 80% and the precision is 75%, what is the TPR?
5. If the recall is 80% and the precision is 75%, what is the FPR?

Note: It may not be possible to answer some/any of these questions without additional information.

Answers

1. The positive class is the presence of the disease.
2. A recall of 80% would mean that 80% of the positive cases were found by the detector (if you submitted the entire population). Alternatively, a recall of 80% means that there is an 80% chance of someone with the disease setting off the detector. The problem with a low recall score is that we would miss people that were unhealthy. If the recall is 80%, we are would not detect 20% of the sick population.

3. A precision of 75% means 75% of the times the detector went off, they were actually positive cases. The problem with a low precision score is spending time having people undergo further screenings or using medication unnecessarily. In this example, 25% of the people we flagged as being sick would have unnecessary followups.
4. TPR is the same as recall; in this case it is 80%.
5. The FPR cannot be found from the information given.

Let's look at the last claim in more detail. There are four numbers in the confusion matrix, but if we double all of them, our metrics don't change (i.e. the things we measure such as precision, recall, etc are normalized to the population). Given that we only have two independent numbers (precision and recall) we cannot expect to recover all the different metrics. We would expect given *three* independent numbers, however, we could. (The number is three because all metrics can be written in terms of the four number TP/FP/TN/FN, but our metrics don't depend on the sum of these four numbers).

How do we *know* that FPR is one of the numbers we cannot recover from precision and recall alone? We could try an algebraic proof, but that isn't that useful for an interview. Instead, let's use the precision and recall given, but assume the percentage of people affected by the disease (the *baserate*) is different in the two cases. By showing we get different numbers for the FPR, we can conclude we cannot calculate the FPR without additional information such as the baserate.

Consider a population with 1200 people, and 5% of the population (60 people) have the disease.

- 80% recall states that we can detect 80% of these people, ie. 48 of the 60 will be detected.
- 75% precision states that 75% of those detected have the disease, ie. the 48 detected people make up 75% of the detections. The classifier classifies 64 people as having the disease, 48 correctly and 16 times incorrectly.
- The FPR is the fraction of the healthy population that set off the detector. There are 1140 healthy people, and 16 of them set off the detector, so the FPR is 16/1140=1.4%.

What if the baserate had been 10% instead? Then we have 120 people out of 1200 that have the disease.

- 80% recall tells us that 96 people with the disease will be detected (80% of the 120 diseased people)
- 75% precision tells us 128 people will set off the detector (96 people is 75% of 128, the other 32 people that set off the detector are healthy)
- The FPR is the fraction of the healthy population (1080) that set off the detector (25% of 128 = 32). The FPR is 32/1080 = 3%.

Since our rate changes depending on the assumed baserate, we can conclude that we don't have enough information from precision and recall alone to calculate the FPR.

Problem 2: Breathalyzer Tests

A breathalyzer registers someone's blood alcohol content to tell if they are "over the limit" or "under the influence" of alcohol. It is typically used at roadside police stops to determine if someone is legally able to drive.

1. What is the positive class?
2. What would a recall of 70% mean?
3. What would a precision of 90% mean?

Answers

1. The postive case would be detecting someone with blood alcohol content (BAC) over the limit.
2. A recall of 70% means there is a 70% chance of someone who has a BAC over the limit of setting off the detector. A low recall would mean that people wiht high BAC are not gettin caught when stopped.
3. A precision of 90% means that 90% of the people who blow a positive result actually have a BAC that is too high. Low precision would mean that a significant fraction of the people who blow a positive result would have a follow-up test (such as a blood test) or a ticket when they were within the legal limit.

If people failing the breathalyzer are immediately ticketed, and have to contest the ticket, then it is preferrable to favor high precision (i.e. ensure that we only give tickets to those that we think are actually guilty, at the expense of letting some guilty people go). If they have to be submitted to a different field test (e.g. blood test) then we can be a little less insistent on precision, as the cost of falsely flagging someone is to pass them on to a secondary test.

This example also shows one of the limitations of precision and recall as measures. If the legal limit is 0.08%, failing to flag someone that has 0.0805% counts as much against recall as failing to flag someone with a 0.10% BAC. Likewise, flagging someone with a 0.0795% counts as much against precision as flagging someone with 0.03%. In a real application, you would want to emphasize precision close to the limit, but transition to prioritizing recall for people well over the limit.

For a breathalyzer, the input is generally just chemical, and we simply have to pick a threshold. If we are looking at a machine learning algorithm that has access to demographic information, we should be very careful when discretizing a continuous variable into two categories and treating mild infractions the same as serious ones.

Problem 3: Should we approve a loan?

We are looking to develop a machine learning algorithm to predict whether someone will pay a loan back or not.

1. What is the positive class?
2. What would a recall of 75% mean?
3. What would a precision of 85% mean?

Answers

1. The positive class are the borrowers that pay back the loans.
2. 75% recall means that 75% of the borrowers that would pay back the loan are approved by our system. We miss 25% of people that would have paid us back by rejecting them. In general, the problem with a low recall is that we are rejecting customers who we would have paid us back (and for whom we would have made interest).
3. 85% precision means that of all the loans we approve, 85% pay us back. The remaining 15% of approved loans go into default. The problem with a low precision is that we are approving loans that are defaulting.

In this example we would generally prefer to emphasize precision over recall, as approving a bad loan (and losing the capital investment) is more costly than missing out on the profit we could make from a good loan.

We have to be a little careful here too, as we are binarizing a continuous variable: there is a difference between someone who defaults after paying 1 year of a 5 year loan, and someone who defaults after paying 4 years of a 5 year loan. A product manager might expect you to look at the expected loss of a customer and threshold on that as a more useful (and nuanced) metric, rather than precision.

Problem 4: Should we unlock a phone?

We are building a facial recognition algorithm to allow people to unlock their phone. If the phone recognizes the person as the authorized user, it will unlock the phone. If it doesn't recognize the user, it will prompt them to try again or try an alternative method (such as a passphrase).

1. What is the positive class?
2. What would a recall of 80% mean?
3. What would a precision of 70% mean?

Answers

Problem 5: Detect malicious programs

When running a program for the first time, we are running some information about the program (such as where it was downloaded, size of the executable, etc) through a classifier. If the program is deemed safe, it will run. If it is deemed unsafe, the user will be prompted to confirm that the program is safe before running.

1. What is the positive class?
2. What does 85% recall mean?
3. What does 75% precision mean?

Answers

Problem 6: de-duplicate records

You are writing a deduplication algorithm. Its goal is to flag entries in your database that are duplicates of existing records. It is more complicated than checking if two records are identical (Which is an easy problem), but instead tries to assess if differences are meaningful. For example:

- The names *Jane Smith*, *Ms Jane Smith*, and *Jane J. Smith*, may refer to the same person
- The addresses *101 Main St Apt 101*, *101 Main St, Apt 101*, *101 Main Street #101* may all refer to the same street address

Records that our algorithm detects as duplicates will be reviewed, and if we are sure they are duplicates, they are removed.

1. What is the positive class?
2. What does 85% recall mean?
3. What does 75% precision mean?

Answers

Summary

When doing interview practice (and in actual interviews) you should translate from the more abstract "positive class" and "negative class" to describe the meaning of precision and recall in the context of the problem you are trying to solve. The difference between precision and recall often trips up people when learning data science; they are nearly incomprehensible when talking to most executives.

You should also have an argument for whether you should be optimizing for precision or recall. In a realistic problem, you shouldn't be optimizing for one or the other -- rather you should look at the *tradeoff* between precision and recall, and pick the best tradeoff for the problem at hand.

The purpose of this article was to provide practice for typical interview questions for data science and machine learning engineer positions. The article *These are't the metrics you're*

looking for looks at the problems that precision and recall have. These problems will be more important when you are actually on the job, or when interviewing for analyst positions.

Related Articles

- These aren't the metrics you're looking for
- Bad names in Classification Problems
- What is a ROC curve?

Metrics

Damien martin

I am a data scientist with an interest in what drives the world. Background in Physics, Math, and Computer Science. Interested in Algorithms, Games, Books, Music, and Martial Arts. That is, when I am not off taking pictures somewhere!

USA Website

Keeping Notebooks Clean

Jupyter notebooks allow for quick experimentation and exploration, but can encourage...

The Bad Names In Classification Problems

There are a proliferation of different metrics in classification problems: accuracy,...