Kirill Bondarenko

Follow 40 Followers About

Precision and recall in recommender systems. And some metrics stuff.

Kirill Bondarenko Feb 23, 2019 · 12 min read

Hello everyone! This story is about how to define quality in recommender systems.

Main topics for discussion are:

1. What metrics can be used in recommender systems to analyze **quality of the given recommendations** ?
2. **Precision and Recall**. How to use them ?

Introduction

I want to start from the very beginning. Problem of choosing right metrics for your task is the second paced problem by difficulty after choosing model type. Because you may choose the best model in the world for giving users recommendations, but if you choose wrong metrics to define its quality you will get strange or bad results. Or you will get good results but won't be able to define why are they good and stuck in this narrow place.

I wrote this article because faced with this problem personally in my task. I found a good model, fitted it by my data, got some results, and found there built in precision and recall metrics. After using it I've got **f1** score (harmonic mean of precision and recall) equals 0.99 ! (maximum is 1.0). So, I wondered and was very happy. But when I looked for results from a human point of view, I wondered in a bad way, it was just a set of trash without sense(or just a piece of sense there was, but no more). I tried other

From this part I started my browsing in the web to find answers. I found academic papers with a tonnes of formulas how it's "easy" to define metrics, but I've got no clear view for the idea of calculating metrics. Finally I found few good articles with explanation. You may find links for them at the bottom of my article. But here I want to make an explanation in my manner by my own words with some graphic examples.

Precision and recall

There are a lot of metrics how to define quality of the model. But the main two are precision and recall. Let's look for their definitions in general.

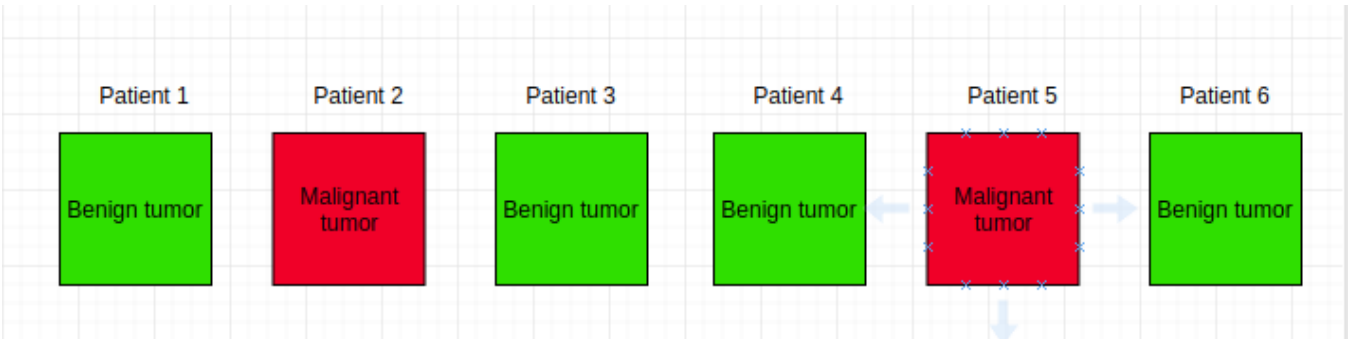
> *In pattern recognition, information retrieval and binary classification, **precision** (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while **recall** (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Both precision and recall are therefore based on an understanding and measure of relevance. — Wikipedia*

It will be easier to look for example in one life situation. If you know definitely what precision and recall mean and can explain it like 2 x 2, you may skip next part till the ***Harmonic means(F-scores)***.

Life example

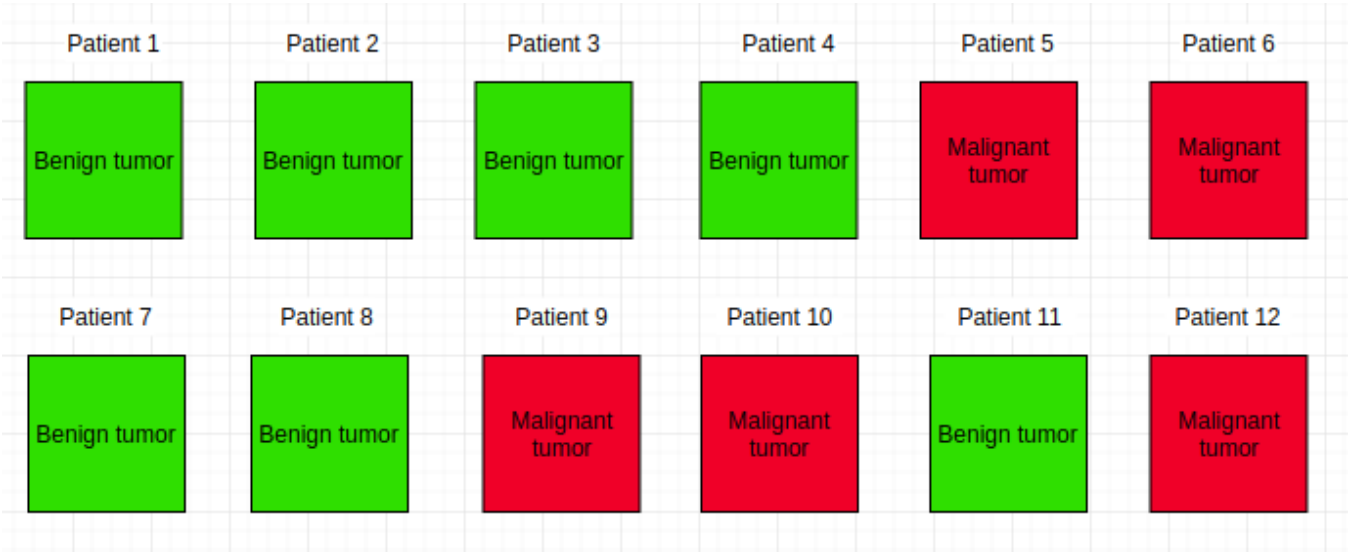
Imagine, you are a doctor in a clinics. You have 12 patients and everyone bring you a fluorography (x-ray picture of the lungs). And your task is to define: does patient have a cancer or no ? To simplify, it's true/false question, where true means patient has a cancer, false means patient doesn't have a cancer (malignant and benign tumors). And your clinics buy for you new software to define it just after scanning of the picture.

And you tell the next to patients. Green color corresponds to benign tumor, red to malignant.



Initial prediction

After this you do some medical research and tests and see the truth.



The truth

First of all we must point what is **relevant** for us ? For sure is to find malignant tumor. Then, we need to define four types of answers: TP, TN, FP, FN, where *TP* = True Positive, *TN* = True Negative, *FP* = False Positive, and *FN* = False Negative.

TP = we predict a malignant tumor and it is actually is

TN = we predict benign tumor and it is actually is

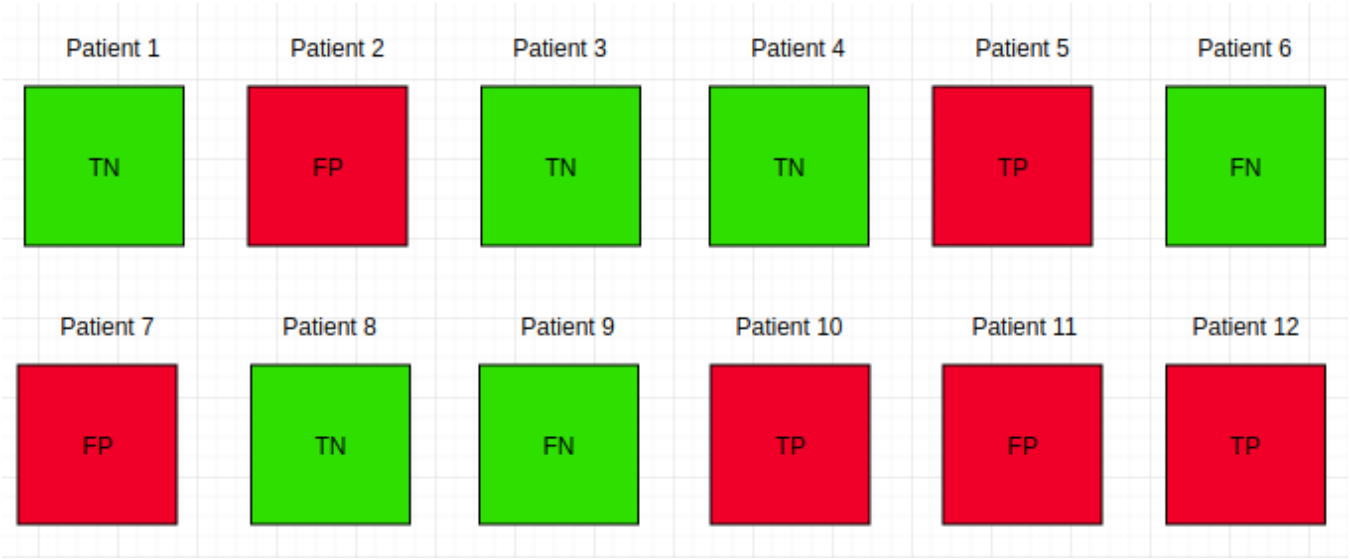
FP = we predict malignant tumor, but it is benign

FN = we predict benign tumor, but it is malignant

So, our aim is to define right TP and TN (perfect model). But we live in a real world and here are no perfect things (only in subjective view, but we are doing science). How are important for us FP and FN ? FP means that probably patient will go to the court and make a complaint. It is not good, but patient will live and it is the most important. FN is the worst result. Patient has a cancer and we told that he/she doesn't.

Precision = TP / (TP + FP) . Characterize model ability to make relevant for us predictions right. It is the ability to make relevant predictions (find malignant tumors).

Recall = TP / (TP+FN) . Describes the part of relevant predictions from all predictions. What part of true malignant tumors have we found ? Best is 1.



Relevant terms

Totals: TP = 3, TN = 4, FN =2, FP = 3

Accuracy = (3 + 4) / (3 + 4 + 2 + 3) = 7/ 12 = 0.58 = > 58 % accuracy

Precision = 3/ 3+ 3= 3/6 = 0.5 => 50 % precision

Recall = 3/ 3+2 = 3/5 = 0.6 => 60% recall

What do these three metrics mean ?

Accuracy 67% mean if we have 100 pictures of lungs , 67 of them will be predicted right in classification benign/malignant and 33 wrong. I think it is a bad result, but not terrible. So, accuracy does not tell all the truth. Let's see deeper. Precision 50% means if we have 100 pictures of true malignant tumors in lungs, only 50 of them will be retrieved. This is really bad result. And finally 60% recall means that we found only 60% from our true malignant tumors. Better then nothing, but model like this is inappropriate for clinics where we need recall 100%.

precision/recall metrics has one hidden surprise. If we increase one, another decreases and vice versa. For example if we point recall as the main quality metrics we may loose precision. But precision must move to 1 too. What we should do ?

If you think about average between precision and recall, you are **almost** right.

Harmonic means (F — scores)

If we have two values A and B, the traditional mean is (sum(A)+sum(B))/2.

Our values are **precision** and **recall**, correspondent to 0.5 and 0.6 . Traditional mean is: $(0.5+0.6)/2 = 0.55$. Is it good or bad we can define after answering one question:

> *Precision and recall are equal for us ?*

We need to imply harmonic mean, when we pay attention to both values in some proportion.

A general view for harmonic mean is next:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}.$$

Harmonic mean or F — score in ML

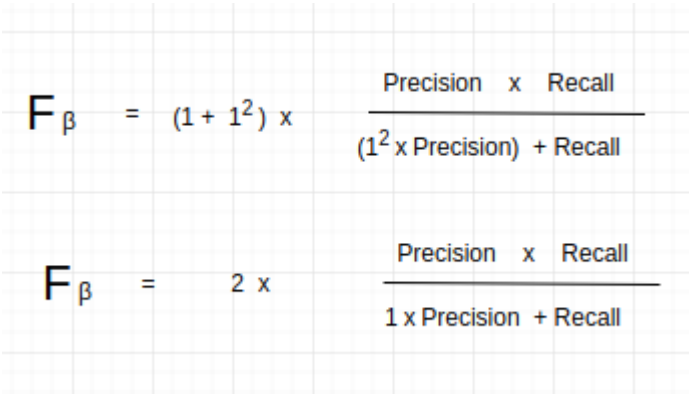
Where **β** — is used in mathematics to indicate when a variable can be entered. The term F 2 score will be used when we want to twice the weight is given to recall opposite to precision. When we twice a weight to precision, an F 0.5 score is used.

F 0.5: When we state that **precision is more important than recall**

$$F_\beta = (1 + 0.5^2) \times \frac{\text{Precision} \times \text{Recall}}{(0.5^2 \times \text{Precision}) + \text{Recall}}$$

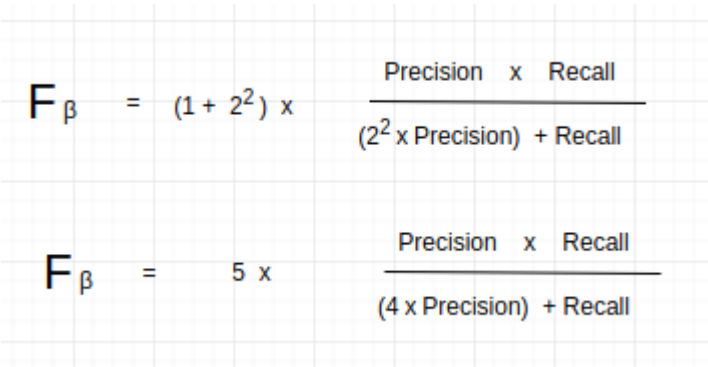
$$F_\beta = (1 + 0.25) \times \frac{\text{Precision} \times \text{Recall}}{(0.25 \times \text{Precision}) + \text{Recall}}$$

F 1: When we state that **recall and precision are both equal for us**



Recall == Precision

F 2: When we state that **recall is more important than precision:**



Recall > Precision

We stated three types of harmonic mean. Now we need to answer the question at the beginning of this part. What does really more important to us ? Precision or recall ? If we say “both” , we may use F 1(which actually equals to traditional mean) and it equals 0.55. Our task is to minimize chance of a false negative results. It means that recall is more important for us than precision. Let’s calculate F 2 for our data.

F 2= 5 x (0.5 x 0.6) / (4 x 0.5 + 0.6) = 1.5/2.6 = 0.58

We got F 2 = 0.58 . It will be our new criteria for model. F 2 score no less then 0.9 ! Sounds great. After this for sure we must increase F 1 score, because model with F 1 score up to 1 is the best model.

How to interpret precision and recall in recommender systems

Let's start from one article. The question was "can we calculate precision and recall in recommendations system ?".

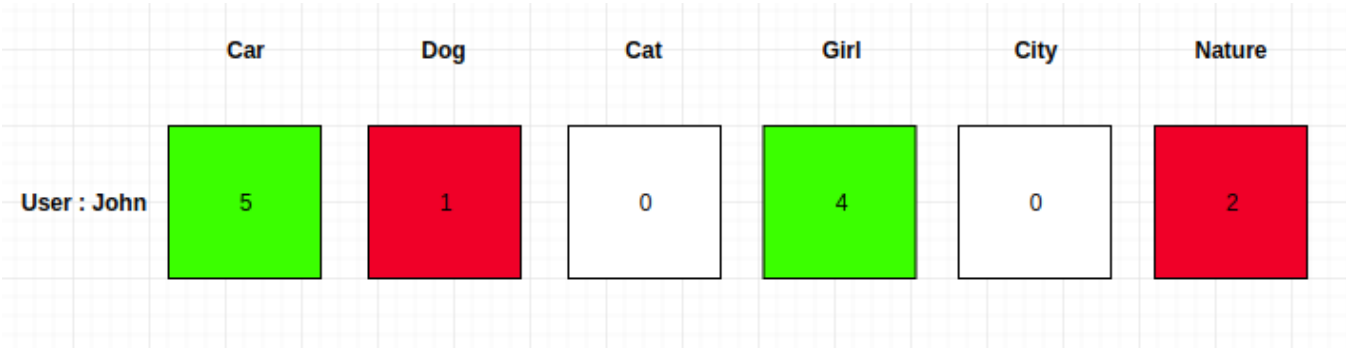
> *In truth,* ***you can't really determine precision and recall in the normal sense*** *— virtually no real recommender application has complete ground truth against which to measure. — Joseph Konstan, Head of GroupLens Research Lab*

Funny, yes ? (I was beating my head into a table while was trying to solve it. It's was not funny for me)

To face this problem right now, we will make a small example.

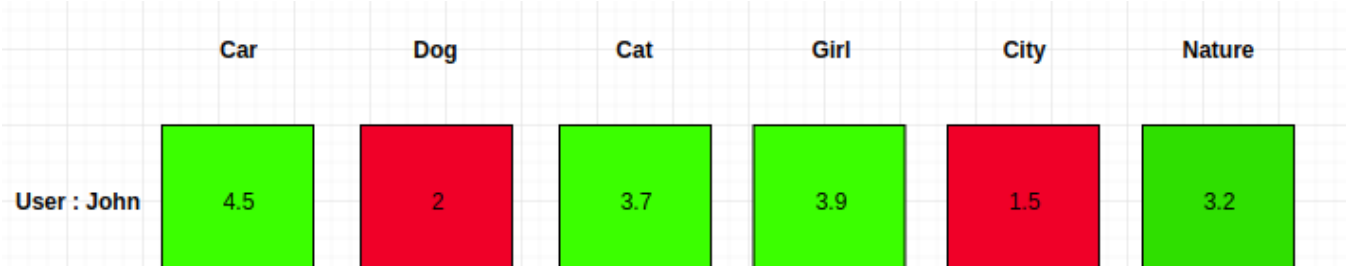
We have a web resource where people post pictures. There are a rating system, where people can rate every picture from 1 to 5 points. Where 1 — user totally dislikes and 5 — totally likes. We have a recommender system (yet we won't pay attention to the model type). This model can predict personal recommendations for users on the basis of their ratings history. It is a classic approach of collaborative filtering. Where we state: ratings from 1 to 3 are bad and from 4 to 5 are good.

Let's take a small example data: **Initial**

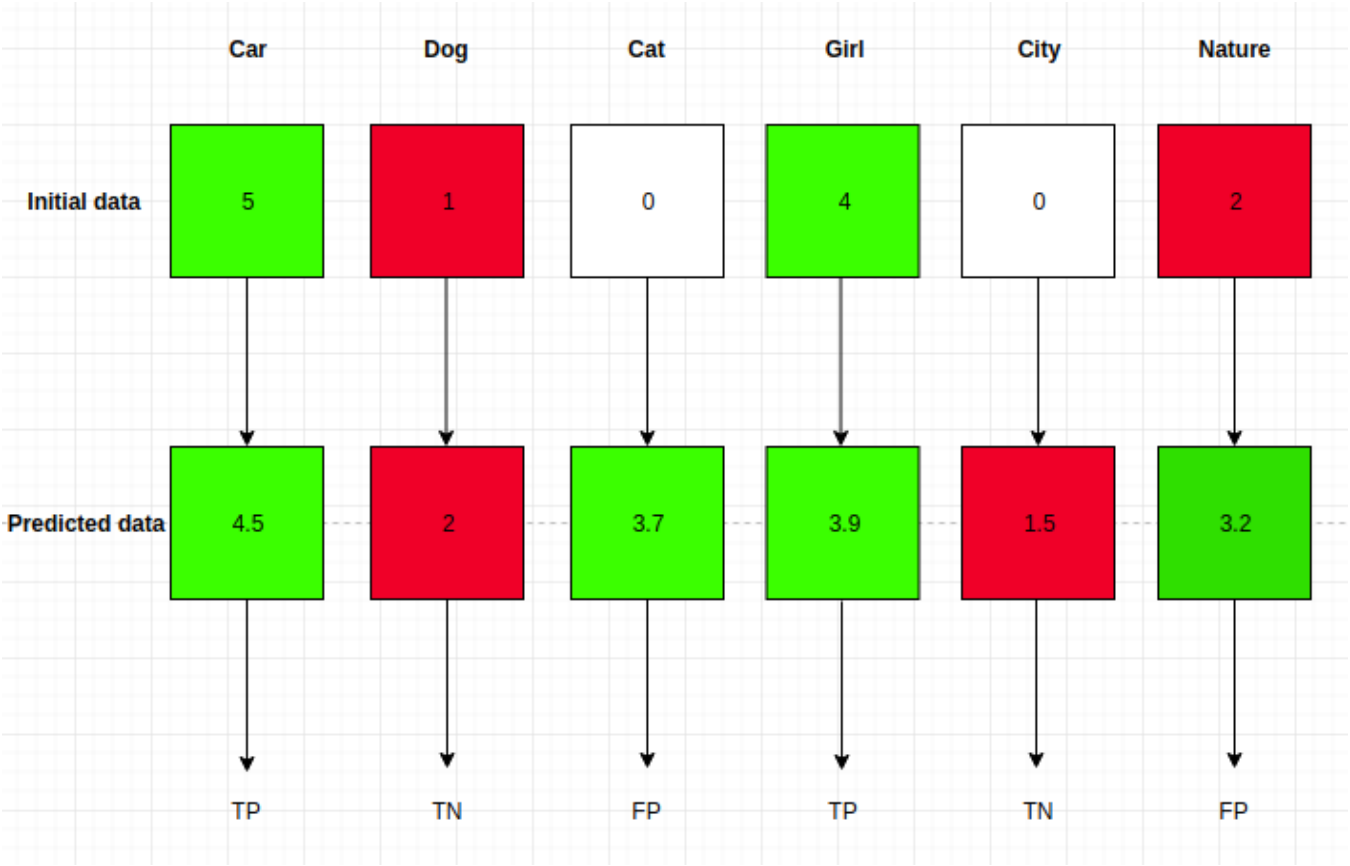


Initial ratings data

Then, our model predict next data: **Predicted**



Let's define classic metrics approach for this task. What is our relevant items(pictures with rating) ? We define, that ratings from 1 to 3 are bad and from 4 to 5 are good. So, out relevant items are pictures with ratings from 4 to 5. Good. Let's do a relevant data.



Relevant answers

TP = 2, TN = 2, FP = 2, FN = 0

And calculate three indicators: accuracy, precision and recall.

Accuracy =(2+2)/6 =0.67 => 67% accuracy

Precision = 2/(2+2) =0.5 =>50% precision

Recall = 2/(2+0) = 1 => 100% recall

Well, we got accuracy 67%, is it good or bad ? Car ,dog and girl pictures were classified right, nature wrong. But here is a one detail.

How to classify 0 — values ? User didn't have an experience with **cat picture**, but we don't know , does it false positive or true positive ? Or how to explain nature

other users told us that this user will like it a bit ?

From all of this stuff we can define next: standard metrics approach doesn't apply for this task. Probably you will hate me after long explanation how to use precision and recall and now I'm telling you that it doesn't fit here ? It's a normal process in data scientist life (remember my words about beating my head into a table ?). We are moving on.

Restructure our metrics approach

Let's relax a bit and talk about psychology. When we want to recommend someone few good bars in a city, do we name all the bars in the city ? — No.

Imagine, your friend asks you: "Hey! Recommend me few good places to drink good beer!". And you start naming all places: "Bar 'A' is good, Bar 'B' is bad, Bar 'C' is bad, Bar 'D' is good , …". Your friend will ask someone adequate. Joking. Funny ? But in this way we tried to evaluate our recommendations in the previous paragraph. For sure you will name only the best bars. Only after this your friend will evaluate your recommendations.

In this way we restructure our precision and recall to **precision at k** and **recall at k.** They also named **P@k** and **R@k,** where **k** — is the number of evaluating recommendations, sorted by value (rating).

Before some actions, I want to define main terms next:

Relevant —pictures that user already liked.

Recommended — pictures that are in the predicted list.

We take k = 3 and sort our example:

> ***Precision*** $= (recommended \cap relevant) / recommended$
>
> ***Recall*** $= (recommended \cap relevant) / relevant$

Precision = 2/3 = 0.67

Recall = 2/2 = 1

If we got data like this:

Precision = 2/2 = 1

But here we faced with a new problem. We just copy user preferences in pictures and got no new pictures. Sense of such recommendations is pure.

In this way for beginning we may state:

> *Recall ->1 is always good and should be there. But model with precision 0.67 gives user more useful information than model with precision 1.*

MAP — mean average precision

Main idea is in the state: we do not take to attention user previous experience. We just evaluate recommended items with the same rating threshold = 3.

MAP = sum(ratings of recommended items)/N recommended items

MAP = (1+0.67+0.75+0.67)/4=0.77 = > 77%

I think it looks better. Because in fact we guess Car and Girl like true positives and recommend Cat and Nature. So we don't care about negative results. It's like an

Conclusion for precision and recall:

What we have discovered ? Precision and recall are the most popular and powerful metrics to evaluate quality of **relevant** predictions. So first we need to define relevance and only after this we may calculate metrics. That's all good in standard classification tasks, but recommender systems live their own life.

If user didn't have an experience with some items but model predict them like positive, what we should do ?

5 items, (1,2,3) are TP and (4,5) are FP : P@5 = 3/5 = 0.6 ; R@5 = 3/3 = 1;

First approach: We must state what is more important for us ? If we choose precision, we will concentrate our metrics on the ability how good we are in retrieving relevant items (already good rated) where as bigger precision than lower our ability to give unique and new recommendations (false positives). Only way to increase precision is to increase TP threshold value. In this way we should increase our f 0.5 score = 0.65 (in last example).

Second approach: f-score between MAP@k and Precision@k

7 items, (1,2,5) are TP , (3,4) are TN , (6,7) are FP.

MAP = (1/1 + 2/2 + 3/5 + 4/6 + 5/7) / 5= 3.98 / 5= 0.8

Precision = 3/(3+2) = 0.6

We state that real precision is more important for us than MAP, so we will use f 0.5 = 1.25 * 0.6 * 0.8 / 0.25 * 0.6 + 0.8 = 0.63

In general, it depends on a purpose of recommender system. If we want to make a marketing campaign where we pay attention to real relevant items , we should use P@k ->1 and R@k ->1 => f1 ->1. If we want to recommend new items and we are confident in our model we may use MAP.

Articles and papers used

Google ML course

MAP explanation

Quora

Article of Maher Malaeb

Thank you for reading and good luck in your tasks!

Bondarenko K. machine learning engineer

Machine Learning Programming Precision Recommendation System Quality