

Evaluating Classifiers : Precision & Recall

- They are better measures of the quality when compared to accuracy.
- Thus far, we considered
 - $\text{accuracy} = \frac{\text{\#correct}}{\text{\#total}}$;
 - $\text{classification error} = \frac{\text{\#mistakes}}{\text{\#total}}$;

Imbalanced Classes

- This can be subjected to the majority class issue. In case of a restaurant review that has majority of negative classes, then the predicted output will be negative. The positive review will be hard to extract.
- Binary classifier : Classification Error - 0.5;
- For k classes, error = $1 - 1/k$
 - error = 0.666 for 3 classes; 0.75 for 4 classes.

Task -> Automated marketing campaign

- The restaurant must display positive review in order to boom.
- Website shows 10 sentences from recent reviews

Precision -> Did I (mistakenly) show a negative sentence? (Show only Positive sentence precisely)

Recall -> Did I not show a (great) positive sentence ? (Show all the positive review from the total)

Precision

Fraction of positive predictions that are actually positive

- It is the fraction of positive predictions that are actually positive.
- Consider the below, the algorithm predicts that six of the sentences are positive, but in reality only 4 are positive.
- Thus, precision = $4 / 6$;

What fraction of the positive predictions are correct?

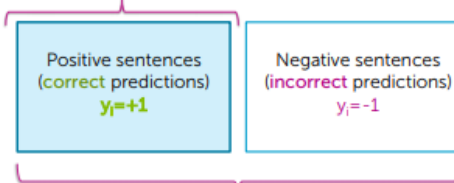
Sentences predicted to be positive: $\hat{y}_i = +1$

Easily best sushi in Seattle.	✓
The seaweed salad was just OK, vegetable salad was just ordinary.	✗
I like the interior decoration and the blackboard menu on the wall.	✓
The service is somewhat hectic.	✗
The sushi was amazing, and the rice is just outstanding.	✓
All the sushi was delicious.	✓

Only 4 out of 6 sentences predicted to be **positive** are actually **positive**

Precision: Fraction of positive predictions that are actually positive

Subset of positive predictions that are actually positive



All sentences predicted to be positive $\hat{y}_i=+1$

Precision - Formula

- Fraction of positive predictions that are correct

$$\text{precision} = \frac{\# \text{ true positives}}{\# \text{ true positives} + \# \text{ false positives}}$$

- Best possible value : 1.0
- Worst possible value : 0.0

Why precision is important

Shown on website

Sentences predicted to be positive: $\hat{y}_i=+1$

Easily best sushi in Seattle	✓
The seaweed salad was just OK, vegetable salad was just ordinary	✗
I like the interior decoration and the blackboard menu on the wall	✓
The service is somewhat hectic	✗
The sushi was amazing, and the rice is just outstanding	✓
All the sushi was delicious	✓

2 negative sentences shown to potential customers... ☹

High precision means positive predictions actually likely to be positive!

Types of error: Review

	Predicted label	
	$\hat{y}_i=+1$	$\hat{y}_i=-1$
True label	$y_i=+1$	True Positive False Negative
	$y_i=-1$	False Positive True Negative

Confusion matrix for sentiment analysis

	Predicted sentiment	
	$\hat{y}_i=+1$	$\hat{y}_i=-1$
True sentiment	$y_i=+1$	+1 sentence +1 prediction -1 prediction → missed a sentence
	$y_i=-1$	-1 sentence +1 prediction -1 prediction → showed bad review on website!! ☹

$$\text{Precision} = \# \text{ true positives} / (\# \text{ true positives} + \# \text{ false positives})$$

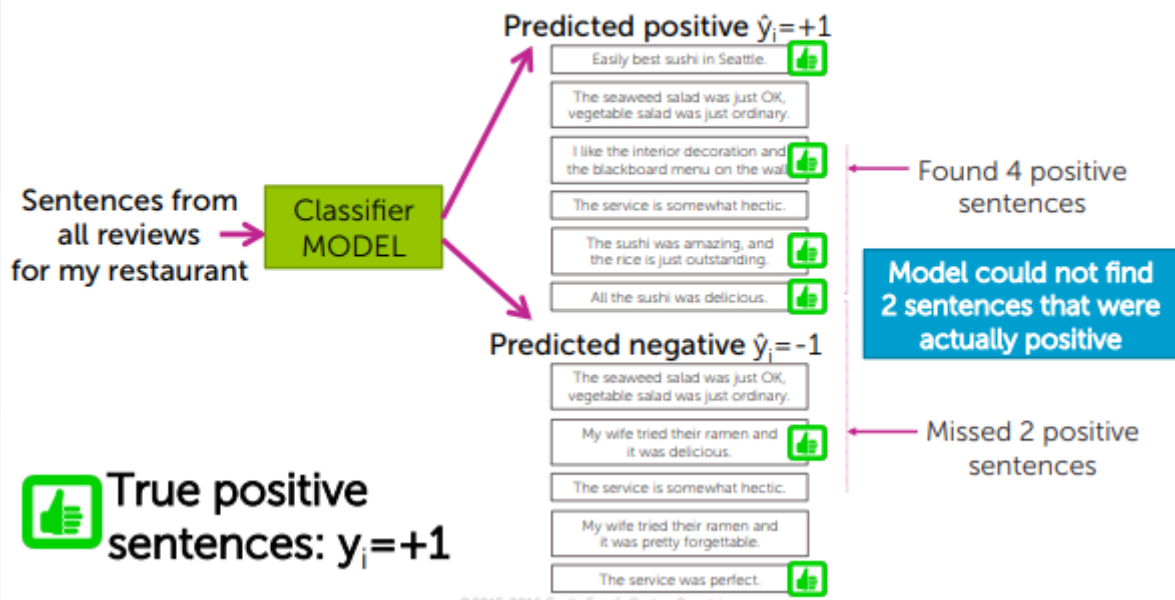
- Best possible value = 1.0
- Worst possible value = 0.0

Recall

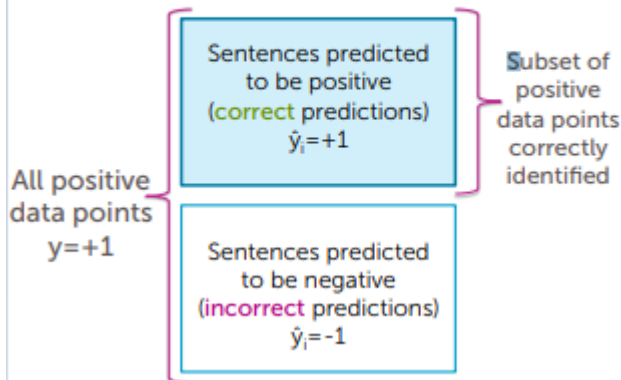
Fraction of positive data predicted to be positive.

- Given a dataset with reviews that's feed to a classifier model and the model predicts.
- Say the model predicts 6 positive and 4 negative.
- But the true label among the positive predictions are 4 while there are 2 true labels lost among the negative predictions.
- Thus the model has lost two positive reviews.

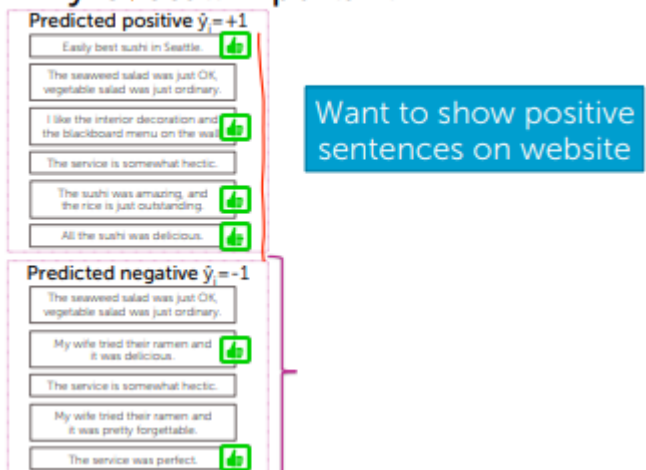
Did I find all the positive sentences?



Recall: Fraction of positive data predicted to be positive



Why is recall important?



Recall - Formula

- Fraction of positive data points correctly classified

$$\text{Recall} = \frac{\# \text{ true positives}}{\# \text{ true positives} + \# \text{ false negatives}}$$

- Best possible value : 1.0
- Worst possible value : 0.0

2 positive sentences not shown to potential customers... ☹️

High recall means positive data points are very likely to be discovered!

©2015-2016, Facebook Inc. & Facebook Foundation

Machine Learning Specialization

$$\text{Recall} = \# \text{ true positives} / (\# \text{ true positives} + \# \text{ false negatives})$$

The precision-recall tradeoff

Precision-recall extremes :

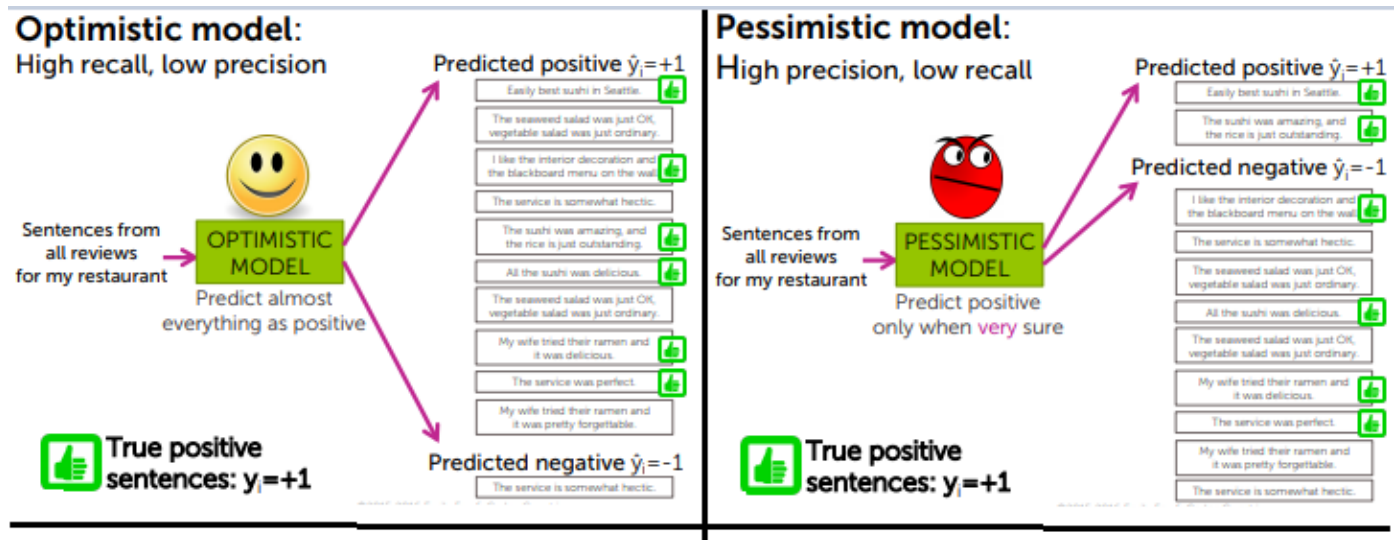
Optimistic Model - High recall, low precision.

- Mostly predicts the data to be positive.
- Hence most of the positive true label will be predicted positive - high recall.
- Since most reviews are predicted positive, many negatives can also be categorized as positives - low precision.

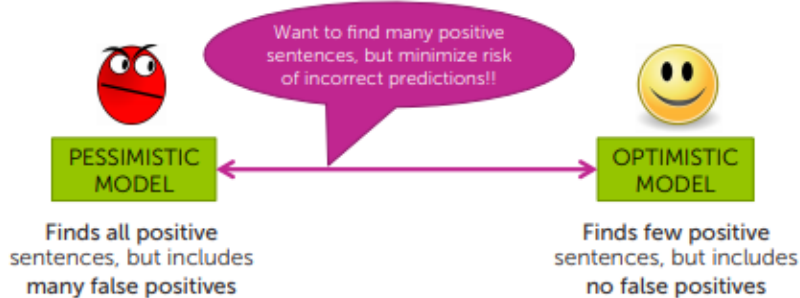
Pessimistic Model - High Precision, low recall.

- Mostly predicts the data negative.
- Hence few records that are predicted to be positive are positive actually, therefore high precision.
- Since most reviews are predicted negative, many positive records will also be marked negative. Therefore low recall.

Therefore require a model that minimizes incorrect predictions.



Balancing precision & recall



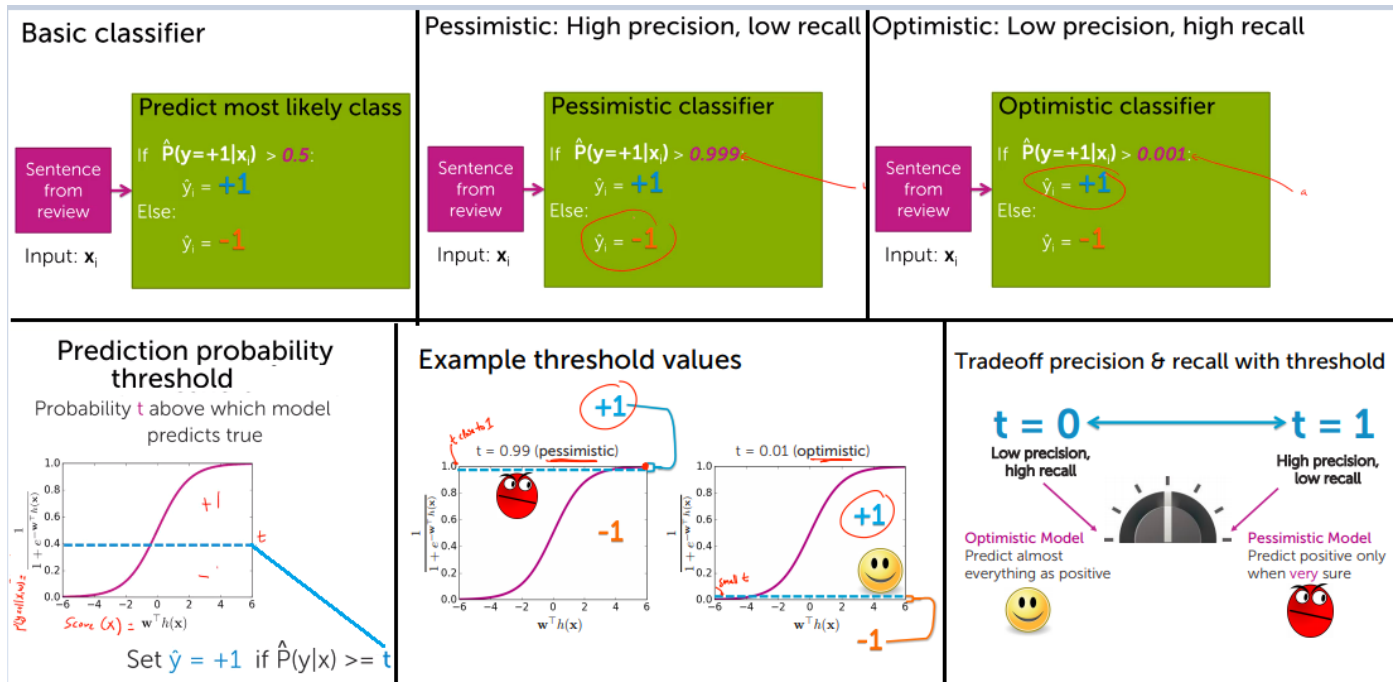
Confidence in the predictions

- In the model, since all reviews cannot be segregated into absolute positives or absolute negatives. A confidence probability is associated with the reviews to indicate and emphasize the sentiment of the review.
- Although the reviews are either +1 (positive) / -1 (negative);
- The confidence probability associated with the reviews can range from 0 to 1.
- This probability can be used to tradeoff precision and recall.

- Thus far, we consider classifiers that considered the threshold prediction probability = 0.5; below that are negative, while above that are positive.

- Optimistic model -> threshold -> 0.001;
- Pessimistic model -> threshold -> 0.999;

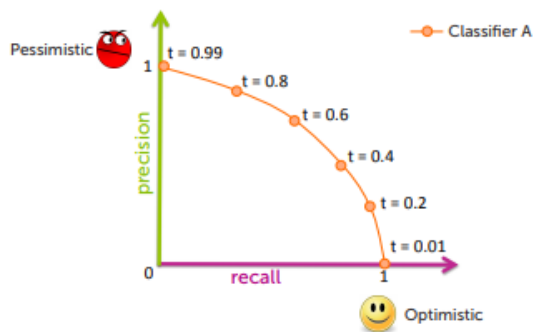
Therefore the tradeoff can be represented by a letter 't' that ranges between 0 & 1.



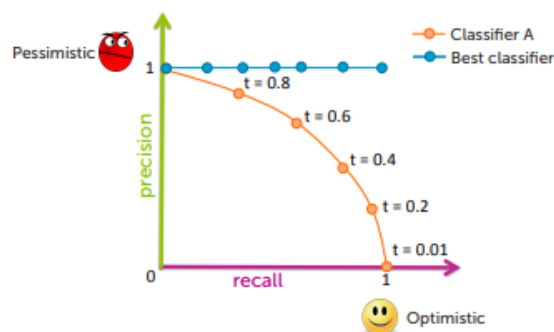
The precision-recall curve

- The best classifier will have a precision 1 irrespective of the recall. But this is an ideal model difficult to achieve in practice.
- For classifiers A, B. B is a better classifier since is closer to the ideal than A.
- For classifiers A, C. It is complicated since there are regions where A is closer to ideal than C and vice-versa.

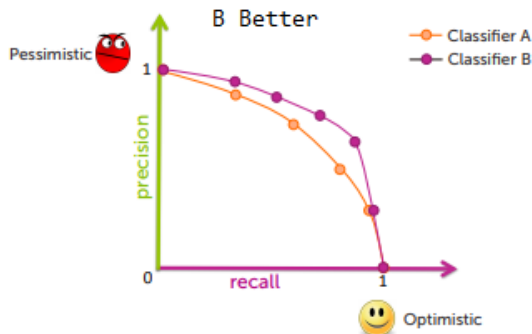
The precision-recall curve



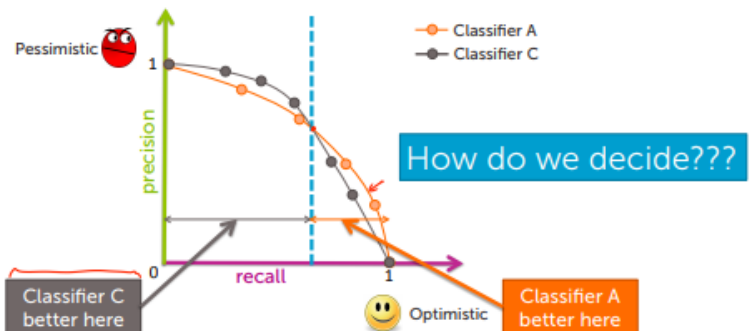
What does the perfect algorithm look like?



Which classifier is better? A or B?



Which classifier is better? A or C?



Compare Algorithms:

- Often reduce the precision-recall to a single number to compare algorithms.
 - F1 measure, area-under-the-curve (AUC), ...
- Precision at k
 - Showing k=5 sentence on websites.
 - Precision (of the given sentences how many are positive = 0.8);

Compare algorithms

- Often, reduce precision-recall to single number to compare algorithms
 - F1 measure, area-under-the-curve (AUC), ...

Precision at k

Showing
k=5 sentences
on website



Sentences model
most sure are positive

- Easy best sushi in Seattle.
- My wife tried their ramen and it was pretty forgettable.
- The sushi was amazing, and the rice is just outstanding.
- All the sushi was delicious.
- The service was perfect.



precision at k = 0.8

Quiz

1. Questions 1 to 5 refer to the following scenario:

Suppose a binary classifier produced the following confusion matrix.

	Predicted Positive	Predicted Negative
Actual Positive	5600	40
Actual Negative	1900	2460

What is the **accuracy** of this classifier? Round your answer to 2 decimal places.

accuracy = total correct / total = 0.81 recall = true positive / true positive + false negative = 0.99 precision = true positive / true positive + false positive = 0.75

2. Refer to the scenario presented in Question 1 to answer the following:

(True/False) This classifier is better than random guessing.

- ☒ True
☐ False
-

recall 0.99 > accuracy 0.81 - random guessing;

3. Refer to the scenario presented in Question 1 to answer the following:

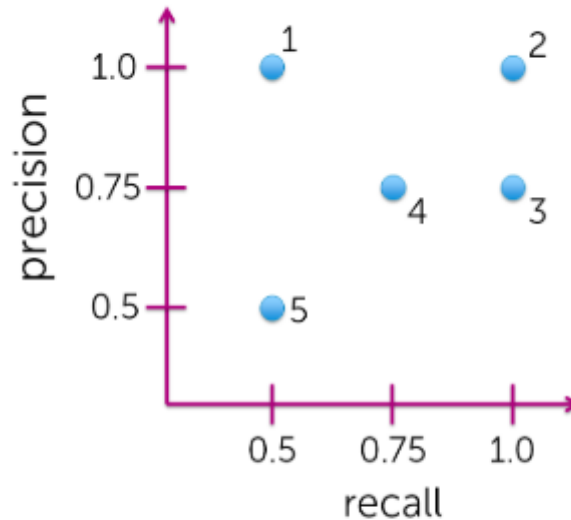
(True/False) This classifier is better than the majority class classifier.

- ☒ True
☐ False

majority classifiers are biased.

4. Refer to the scenario presented in Question 1 to answer the following:

Which of the following points in the precision-recall space corresponds to this classifier?



- ☐ (1)
☐ (2)
☒ (3)
☐ (4)
☐ (5)

Answer

Precision: $5600 / (5600 + 1900) = 0.75$ Recall = $5600 / (5600 + 40) = 0.99$

5. Refer to the scenario presented in Question 1 to answer the following:

Which of the following best describes this classifier?

- ☒ It is optimistic
☐ It is pessimistic
☐ None of the above

- recall > precision -> optimistic

6. Suppose we are fitting a logistic regression model on a dataset where the vast majority of the data points are labeled as positive. To compensate for overfitting to the dominant class, we should

- ☒ Require higher confidence level for positive predictions
☐ Require lower confidence level for positive predictions

More info: <https://www.coursera.org/learn/ml-classification/lecture/IMHs2/trading-off-precision-and-recall>
 (<https://www.coursera.org/learn/ml-classification/lecture/IMHs2/trading-off-precision-and-recall>)

7. It is often the case that false positives and false negatives incur different costs. In situations where false negatives cost much more than false positives, we should
- ☐ Require higher confidence level for positive predictions
- ☒ Require lower confidence level for positive predictions

8. We are interested in reducing the number of false negatives. Which of the following metrics should we primarily look at?
- ☐ Accuracy
- ☐ Precision
- ☒ Recall

9. Suppose we set the threshold for positive predictions at 0.9. What is the lowest score that is classified as positive? Round your answer to 2 decimal places.

2.20

- Class probability \neq score.
- In the context of linear classifier, score is the dot product of coefficients and features.
- Recall that $P(y = +1 | x, w) = \text{sigmoid}(\text{score})$.
- If we want $P(y=+1|x, w)$ to be greater than 0.9, how large should the score be?

$$\begin{aligned}\frac{1}{1+e^{-\text{score}}} &= 0.9 \\ \Rightarrow 0.9 + 0.9e^{-\text{score}} &= 1 \\ \Rightarrow \frac{0.1}{0.9} &= e^{-\text{score}} \\ \Rightarrow \ln\left(\frac{0.1}{0.9}\right) &= \ln(e^{-\text{score}}) \\ \Rightarrow \text{score} &= 2.20\end{aligned}$$