Evaluating classifiers: Precision & Recall

Emily Fox & Carlos Guestrin
Machine Learning Specialization
University of Washington

Using reviews to promote my restaurant

Goal: increase
guests by 30%

Need an automated, "authentic" marketing campaign

Reviews

Great quotes

"Easily best sushi in Seattle."

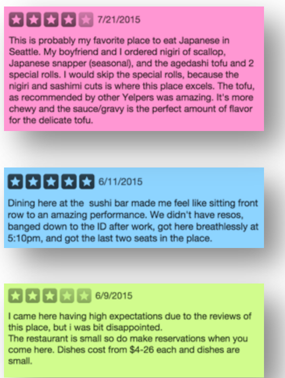
Great spokespeople

How do I find sentences with positive sentiment?

All reviews
for my restaurant

Intelligent restaurant review system

All reviews for restaurant



Break all reviews into sentences

The seaweed salad was just OK, vegetable salad was just ordinary.

I like the interior decoration and the blackboard menu on the wall.

All the sushi was delicious.

My wife tried their ramen and it was pretty forgettable.

The sushi was amazing, and the rice is just outstanding.

The service is somewhat hectic.

Easily best sushi in Seattle.

Sentiment classifier

Input $\mathbf{x}_i$: Easily best sushi in Seattle.

Output: $\hat{y}_i$
Predicted sentiment

Use the sentiment classifier model!

Show sentences with "positive" prediction on website

Sentences from all reviews for my restaurant

The seaweed salad was just OK, vegetable salad was just ordinary.

I like the interior decoration and the blackboard menu on the wall.

All the sushi was delicious.

My wife tried their ramen and it was pretty forgettable.

The sushi was amazing, and the rice is just outstanding.

The service is somewhat hectic.

Easily best sushi in Seattle.

Classifier MODEL

Sentences predicted to be positive

$\hat{y} = +1$

- Easily best sushi in Seattle.
- I like the interior decoration and the blackboard menu on the wall.
- All the sushi was delicious.
- The sushi was amazing, and the rice is just outstanding.

Sentences predicted to be negative

$\hat{y} = -1$

- The seaweed salad was just OK, vegetable salad was just ordinary.
- My wife tried their ramen and it was pretty forgettable.
- The service is somewhat hectic.

What does it mean for a classifier to be good?

Previously, we asked the question:

"What is good accuracy?"

We explored accuracy of random classifier as baseline

- For binary classification:
  - Half the time, you'll get it right! (on average)
    - ➔ classification error = 0.5
- For k classes, error= 1 - 1/k
  - error = 0.666 for 3 classes, 0.75 for 4 classes,…

At the very, very, very least,
you should healthily beat random…
Otherwise, it's (usually) pointless…

We explored the pitfalls of imbalanced problems: *Is 90% accuracy good? Depends ...*

90% of sentences are negative!

90% accuracy by predicting every sentence is negative!!!

Amazing "performance" but not useful for me right now!

Automated marketing campaign cares about something else…

Website shows 10 sentences from recent reviews

PRECISION

Did I (mistakenly) show a negative sentence???

RECALL

Did I not show a (great) positive sentence???

Accuracy doesn't capture these issues well…

Precision:
Fraction of positive predictions
that are actually positive

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

Positive sentences
(correct predictions)
$y_i=+1$

Negative sentences
(incorrect predictions)
$y_i=-1$

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

Types of error: *Review*

Predicted label

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

Confusion matrix for sentiment analysis

Predicted sentiment

True sentiment	$\hat{y}_i=+1$	$\hat{y}_i=-1$
$y_i=+1$	+1 sentence +1 prediction	+1 sentence -1 prediction
$y_i=-1$	-1 sentence +1 prediction	-1 sentence -1 prediction

missed a sentence

showed bad review on website!!

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

Example: Calculating precision

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.	✓

4 correct

2 mistakes

$$\text{precision} = \frac{4}{4+2}$$

$$= 2/3$$

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!

Recall:
Fraction of positive data predicted to be positive

Did I find all the positive sentences?

Sentences from all reviews for my restaurant → Classifier MODEL

Predicted 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.

Predicted negative $\hat{y}_i=-1$

- The seaweed salad was just OK, vegetable salad was just ordinary.
- My wife tried their ramen and it was delicious.
- The service is somewhat hectic.
- My wife tried their ramen and it was pretty forgettable.
- The service was perfect.

True positive sentences: $y_i=+1$

What fraction of positive sentences were missed out?

Predicted 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. 👍

Found 4 positive sentences

Predicted negative $\hat{y}_i=-1$

- The seaweed salad was just OK, vegetable salad was just ordinary.
- My wife tried their ramen and it was delicious. 👍
- The service is somewhat hectic.
- My wife tried their ramen and it was pretty forgettable.
- The service was perfect. 👍

Missed 2 positive sentences

Model could not find 2 sentences that were actually positive

Recall: Fraction of positive data predicted to be positive

All positive data points
$y=+1$

Sentences predicted to be positive (correct predictions)
$\hat{y}_i=+1$

Subset of positive data points correctly identified

Sentences predicted to be negative (incorrect predictions)
$\hat{y}_i=-1$

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

Why is recall important?

Predicted 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.

Predicted negative $\hat{y}_i=-1$

- The seaweed salad was just OK, vegetable salad was just ordinary.
- My wife tried their ramen and it was delicious.
- The service is somewhat hectic.
- My wife tried their ramen and it was pretty forgettable.
- The service was perfect.

Want to show positive sentences on website

2 positive sentences not shown to potential customers… ☹

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

Precision-recall extremes

Optimistic model:

High recall, low precision

Sentences from all reviews for my restaurant

OPTIMISTIC MODEL

Predict almost everything as positive

Predicted 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.
- The seaweed salad was just OK, vegetable salad was just ordinary.
- My wife tried their ramen and it was delicious.
- The service was perfect.
- My wife tried their ramen and it was pretty forgettable.

Predicted negative $\hat{y}_i=-1$

- The service is somewhat hectic.

True positive sentences: $y_i=+1$

Pessimistic model:

High precision, low recall

Sentences from all reviews for my restaurant

PESSIMISTIC MODEL

Predict positive only when very sure

Predicted positive $\hat{y}_i = +1$

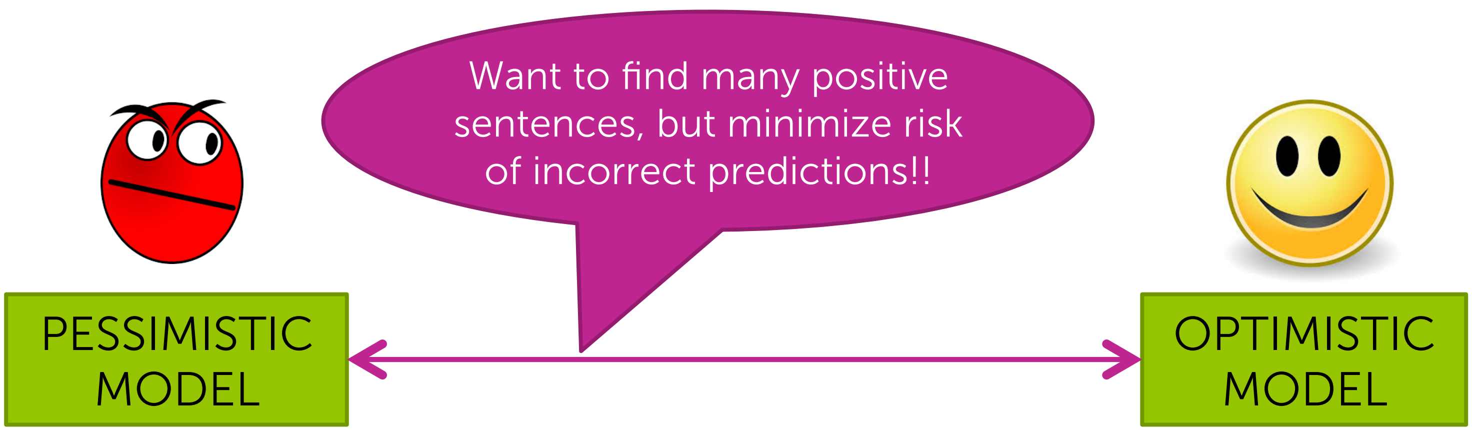
- Easily best sushi in Seattle.
- The sushi was amazing, and the rice is just outstanding.

Predicted negative $\hat{y}_i = -1$

- I like the interior decoration and the blackboard menu on the wall
- The service is somewhat hectic.
- The seaweed salad was just OK, vegetable salad was just ordinary.
- All the sushi was delicious.
- The seaweed salad was just OK, vegetable salad was just ordinary.
- My wife tried their ramen and it was delicious.
- The service was perfect.
- My wife tried their ramen and it was pretty forgettable.
- The service is somewhat hectic.

True positive sentences: $y_i = +1$

Balancing precision & recall



Finds all positive sentences, but includes **many false positives**

Finds few positive sentences, but includes **no false positives**

The description for each model is for the opposite.
Need to swap them to ensure the image is correct.

Tradeoff precision and recall

Can we tradeoff precision & recall?

Low precision,
high recall

High precision,
low recall

Optimistic Model
Predict almost
everything as positive

Pessimistic Model
Predict positive only
when very sure

How confident is your prediction?

"The sushi & everything else were awesome!"

Definite +1

$P(y=+1|\mathbf{x}=$ *"The sushi & everything else were awesome!"* $) = 0.99$

"The sushi was good, the service was OK"

Not sure

$P(y=+1|\mathbf{x}=$ *"The sushi was good, the service was OK"* $) = 0.55$

Can be used to tradeoff precision and recall

Basic classifier

Sentence from review

Input: $\mathbf{x}_i$

Predict most likely class

If $\hat{P}(y=+1|\mathbf{x}_i) > 0.5$:

$\hat{y}_i = +1$

Else:

$\hat{y}_i = -1$

Pessimistic: High precision, low recall

Sentence from review

Input: $\mathbf{x}_i$

Pessimistic classifier

If $\hat{P}(y=+1|\mathbf{x}_i) > 0.999$:

$\hat{y}_i = +1$

Else:

$\hat{y}_i = -1$

very few

Predict positive only when very sure

Optimistic: Low precision, high recall

Sentence from review

Input: $\mathbf{x}_i$

Optimistic classifier

If $\hat{P}(y=+1|\mathbf{x}_i) > 0.001$:

$\hat{y}_i = +1$

Else:

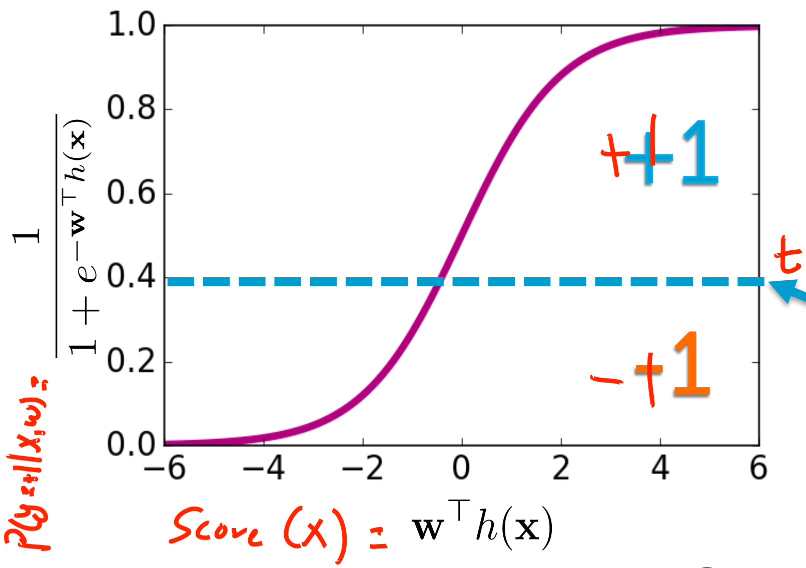
$\hat{y}_i = -1$

almost all

Predict positive almost always

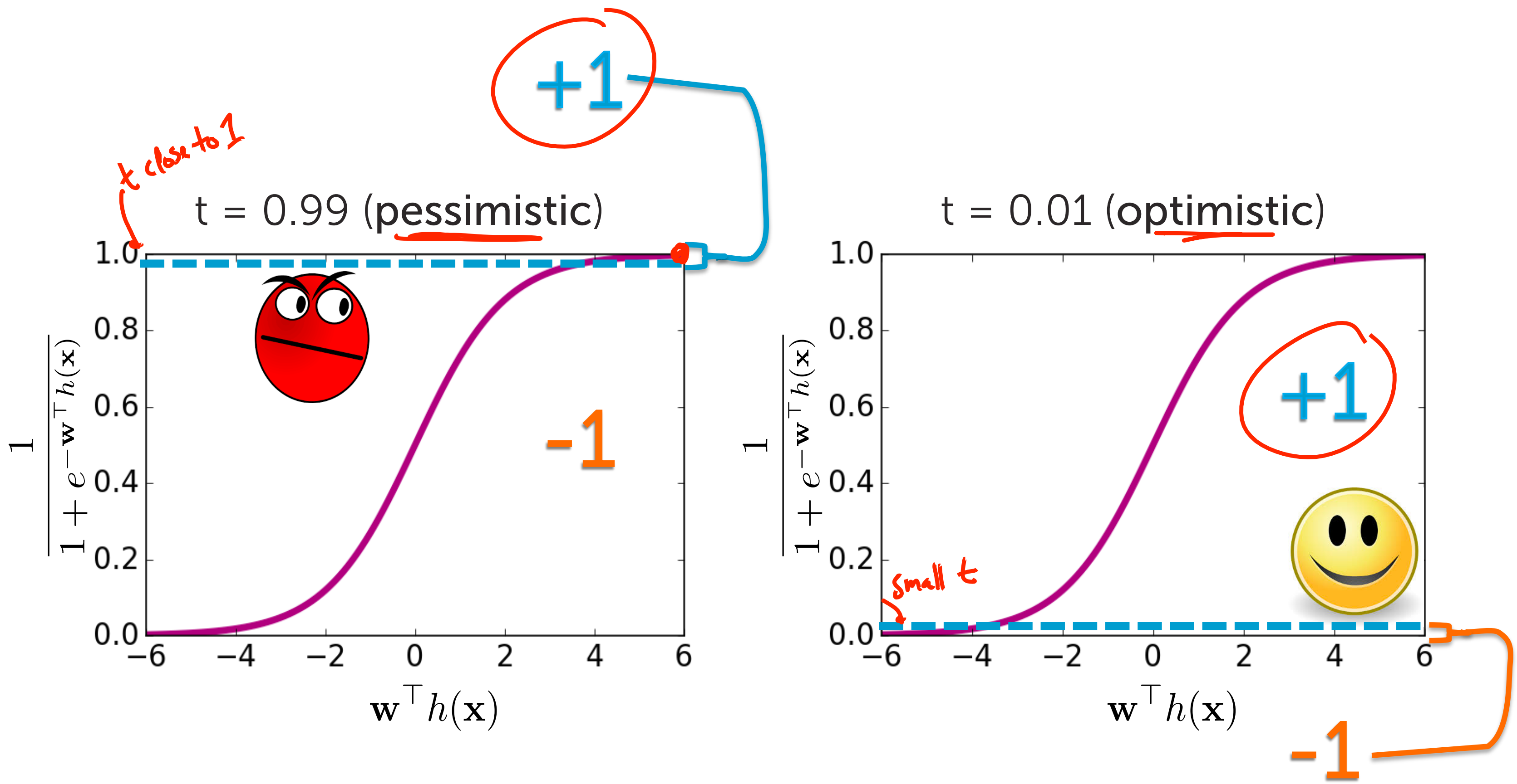
Prediction probability threshold

Probability t above which model predicts true



Set $\hat{y}$ = +1 if $\hat{P}$(y|x) >= t

Example threshold values



Tradeoff precision & recall with threshold

t = 0 ⟷ **t = 1**

Low precision,
high recall

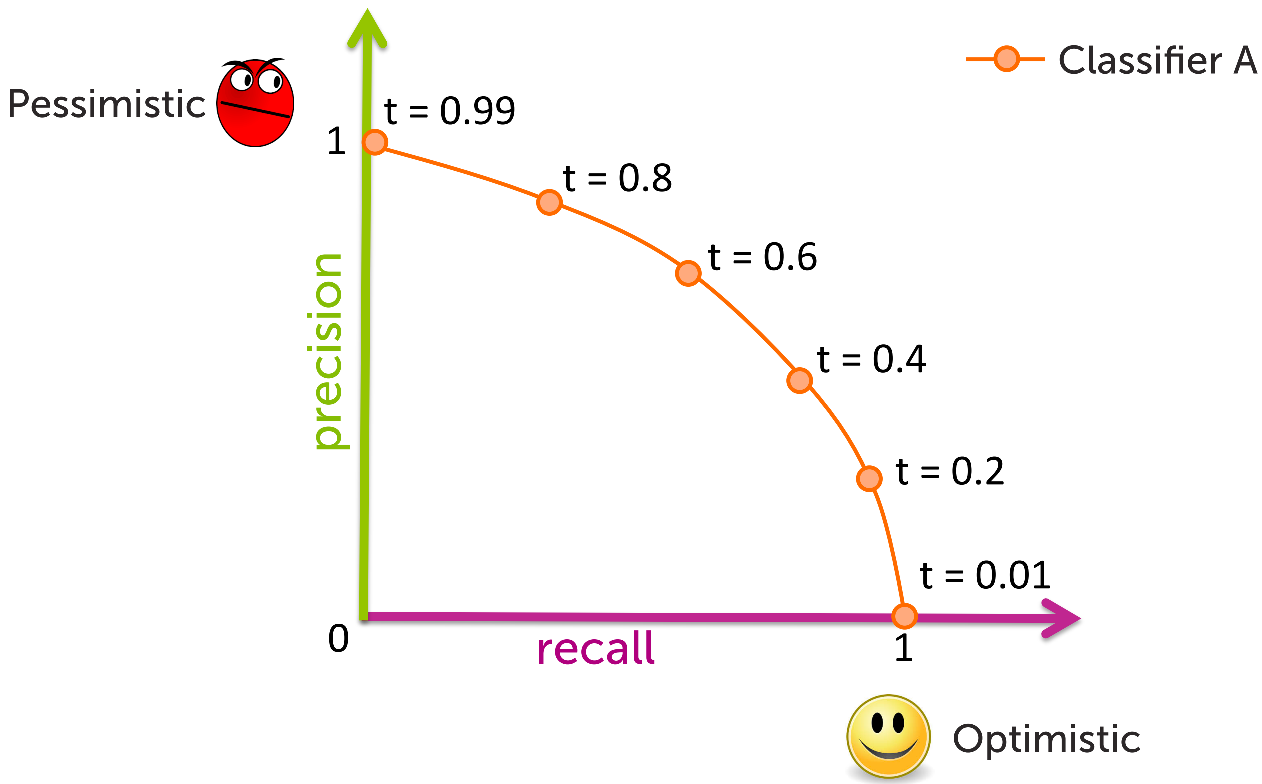
High precision,
low recall

Optimistic Model
Predict almost
everything as positive

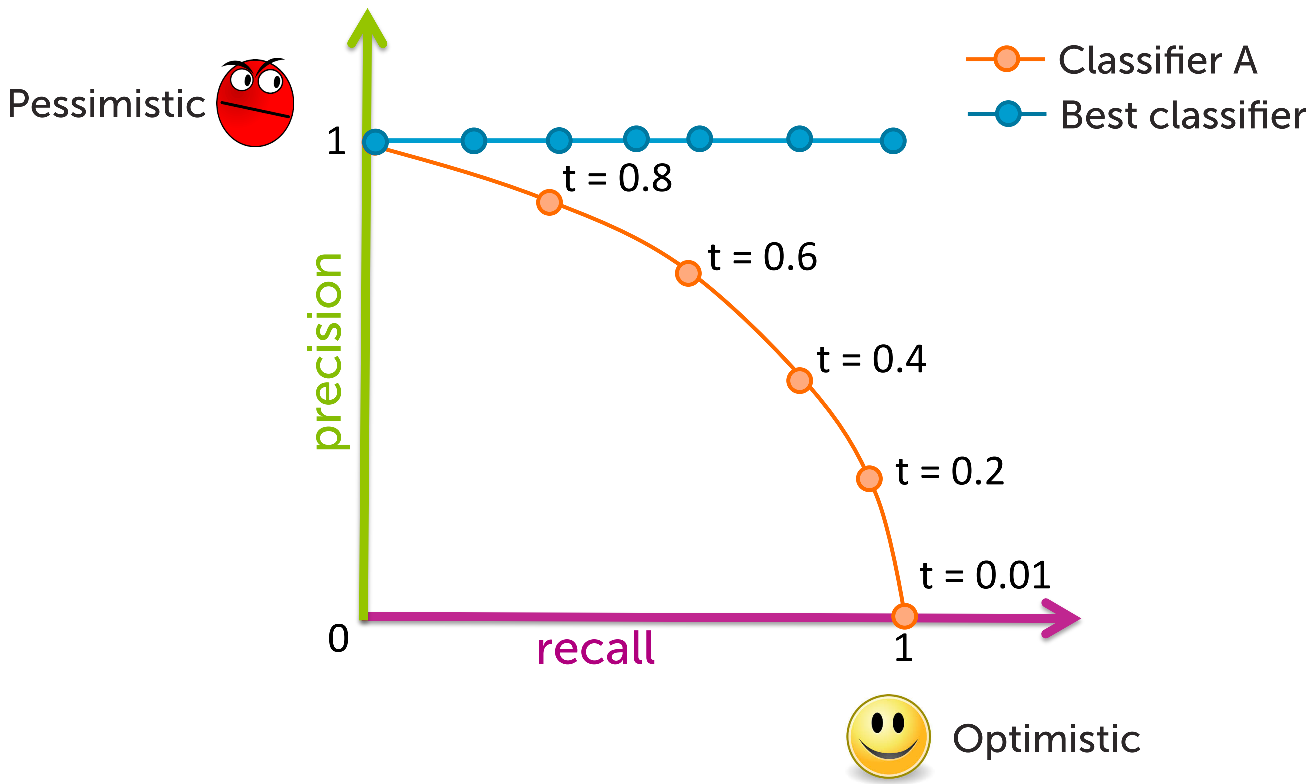
Pessimistic Model
Predict positive only
when very sure

Precision-recall curve

The precision-recall curve

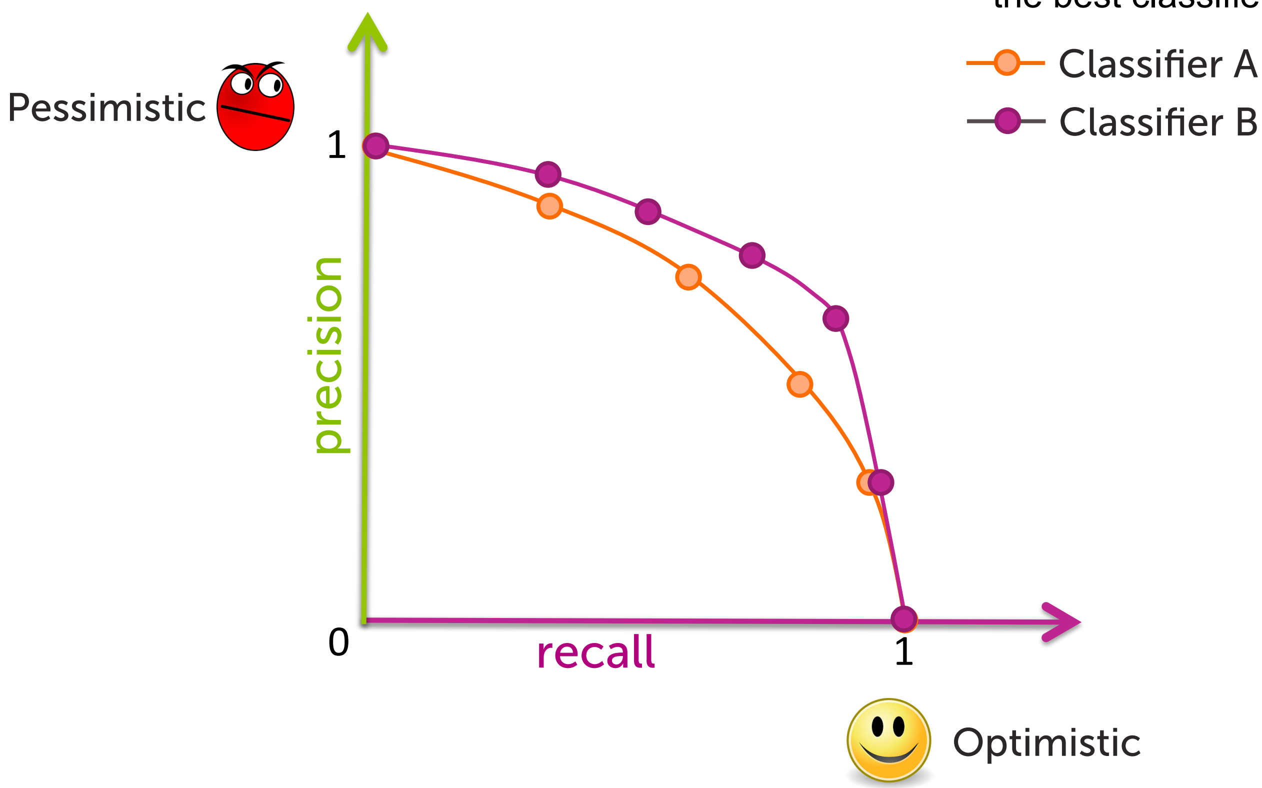


What does the perfect algorithm look like?

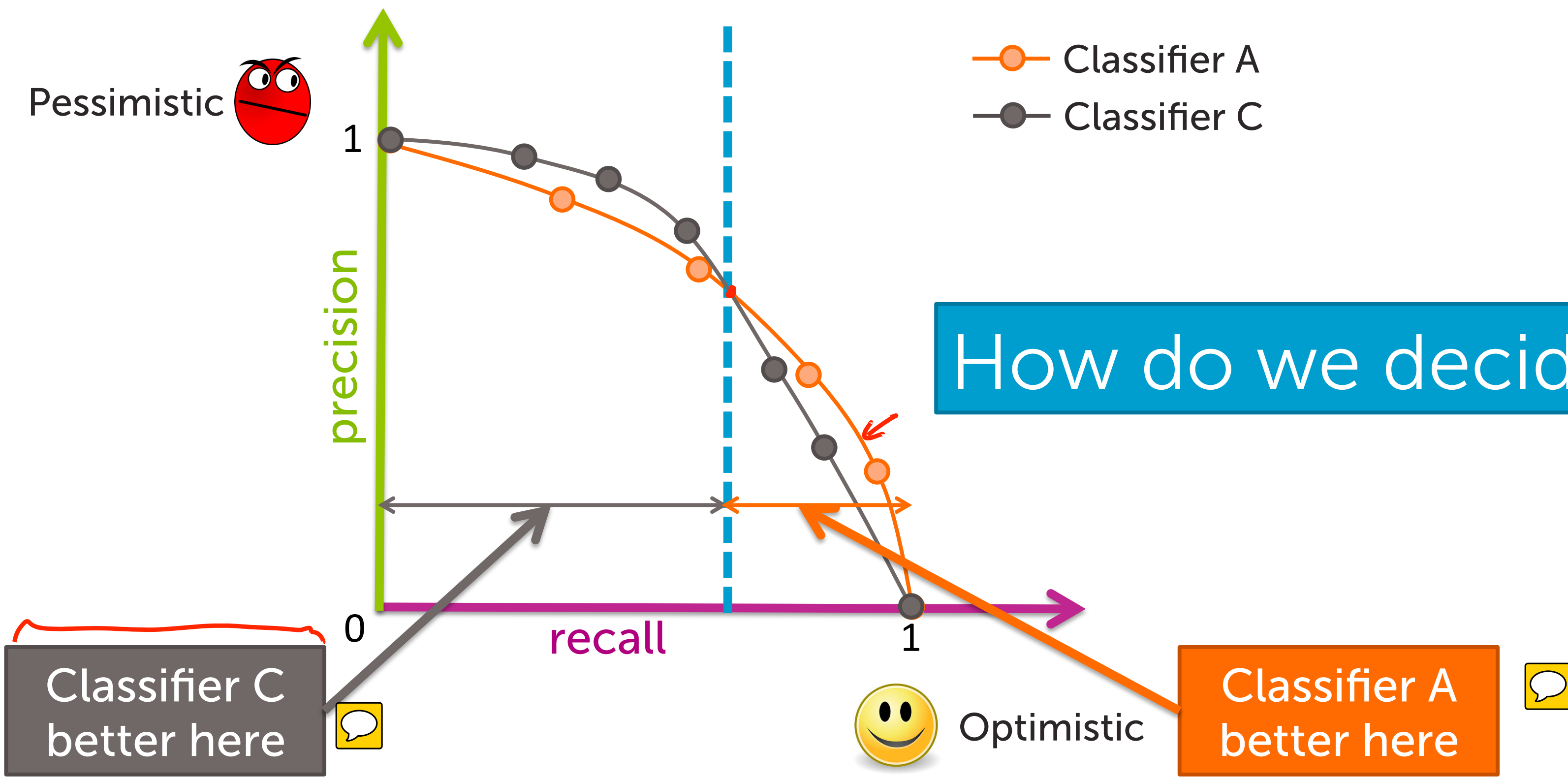


Which classifier is better? A or B?

B, since B is alway higher than A (i.e., closer to the best classifier).



Which classifier is better? A or C?



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

- Easily 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

Summary of precision-recall

What you can do now…

- Classification accuracy/error are not always right metrics
- Precision captures fraction of positive predictions that are correct
- Recall captures fraction of positive data correctly identified by the model
- Trade-off precision & recall by setting probability thresholds
- Plot precision-recall curves.
- Compare models by computing precision at k

Thank you to Dr. Krishna Sridhar

Dr. Krishna Sridhar

Staff Data Scientist, Dato, Inc.