Precision & Recall

Sewoong Oh

CSE/STAT 416
University of Washington

Evaluating a classifier

Example

- Consider a restaurant owner who wants to use ML to promote his business
  - Crawl reviews from Yelp on his restaurant
  - Classify each review as {positive,negative} sentiment
  - Post positive reviews on web

- Need sentiment classifier:

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

Sentence Sentiment Classifier

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

Easily best sushi in Seattle.

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.

- Among many choices of classifiers, logistic regression, decision trees, boosting, etc., which one should we use?

Accuracy

$$\text{Accuracy} = \frac{C_{tp} + C_{tn}}{N}$$

- where True Positive $\boldsymbol{C_{tp}}$ is the number of examples in (test or train) data with $y_i = +1 \text{ and } \hat{y}_i = +1$

- True Negative $\boldsymbol{C_{tn}}$ is the number of examples with $y_i = -1 \text{ and } \hat{y}_i = -1$

- we can use as a baseline a random guess, to get a sense of what is a good accuracy or not

  - for binary classification, a good classifier should at the minimum get better than accuracy 1/2 (as that is what random guess with no training gets)

  - for k-ary classification, the baseline is accuracy 1/k (or equivalently error 1-1/k)

But high accuracy does not always mean good classifier

- If 99% of people do not have cancer, then predicting "no cancer" is 99% accurate, but meaningless

- In the restaurant example, it will choose 10 reviews to show on its website

- A measure of performance that is widely used is **precision** and **recall**

  - Precision: did I (mistakenly) show a negative sentence?
  - Recall: did I miss any (potentially great) positive 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.

Predicted label

		$\hat{y}_i$=+1	$\hat{y}_i$=-1
True label	y_i=+1	True Positive C_{tp}	False Negative C_{fn}
	y_i=-1	False Positive C_{fp}	True Negative C_{tn}

Precision

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

- We care about "how many negative reviews did I include in my list of **n** positively predicted examples?

- precision is designed to answer this question: **n*precision**

$$\text{precision} = \frac{C_{tp}}{C_{tp} + C_{fp}}$$

Recall

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

- We also care about "how many positive reviews did I miss (or include)?"
- Recall is designed to answer this question

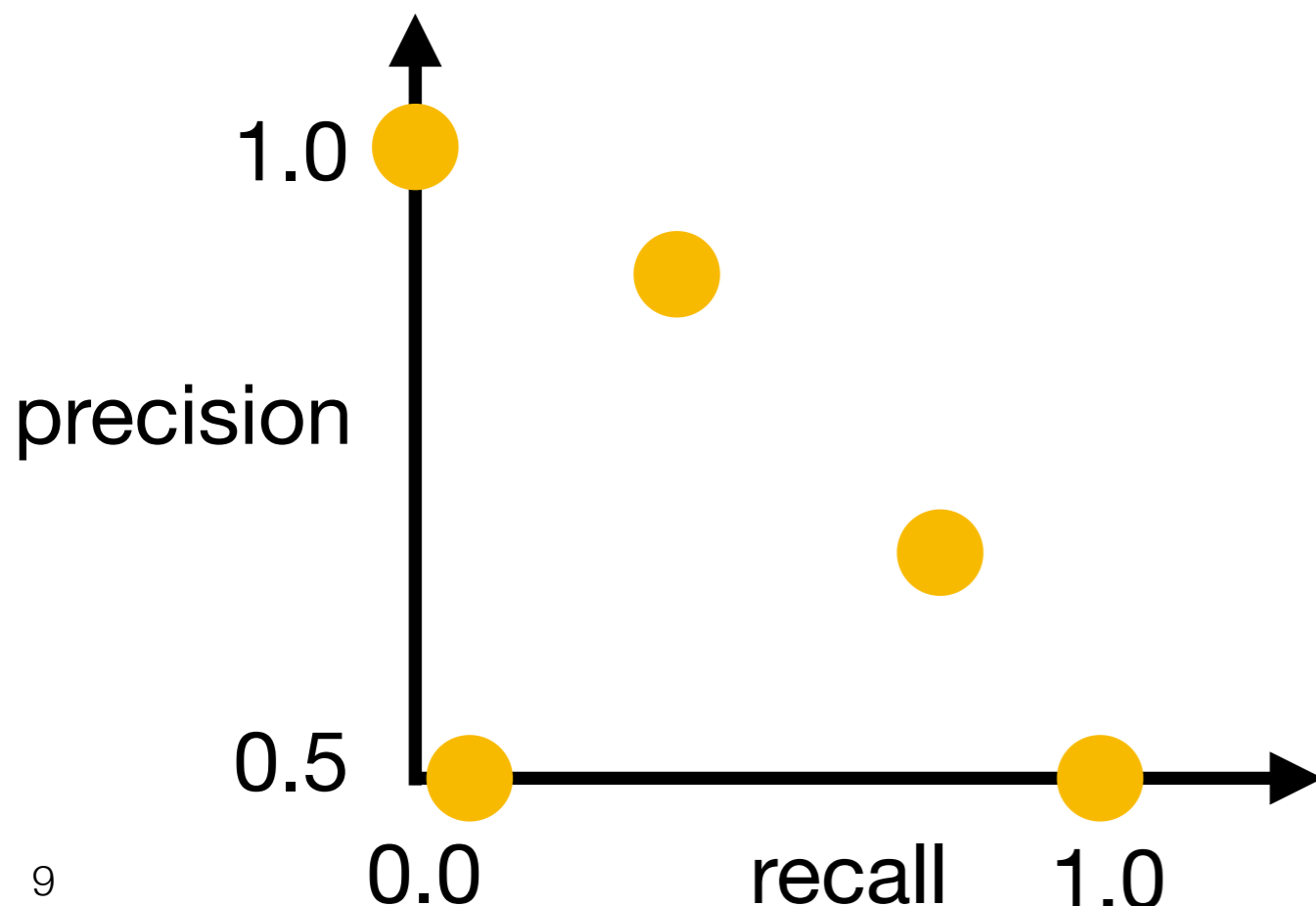$$\text{recall} = \frac{C_{tp}}{C_{tp} + C_{fn}}$$

precision-recall curve

- given training data, each trained classifier achieves a certain (recall,precision) on the test data

- which we can plot in 2-d as below

- for both precision and recall, higher is better

- it is easy to achieve certain points in the graph

- most reasonable models will be incomparable

$$\text{precision} = \frac{C_{tp}}{C_{tp}+C_{fp}} \qquad \text{recall} = \frac{C_{tp}}{C_{tp}+C_{fn}}$$

Predicted label

	$\hat{y}_i=+1$	$\hat{y}_i=-1$
$y_i=+1$	True Positive C_{tp}	False Negative C_{fn}
$y_i=-1$	False Positive C_{fp}	True Negative C_{tn}

True label

precision

1.0

0.5

0.0   recall   1.0

Precision-Recall tradeoff

Want to find many positive sentences, but minimize risk of incorrect predictions!!

PESSIMISTIC MODEL ⟷ **OPTIMISTIC MODEL**

Finds few positive sentences, but includes no false positives

Finds all positive sentences, but includes many false positives

Pessimistic Model

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.	

Plot

precision axis: 1.0, 0.5
recall axis: 0.0, 1.0

Optimistic Model

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

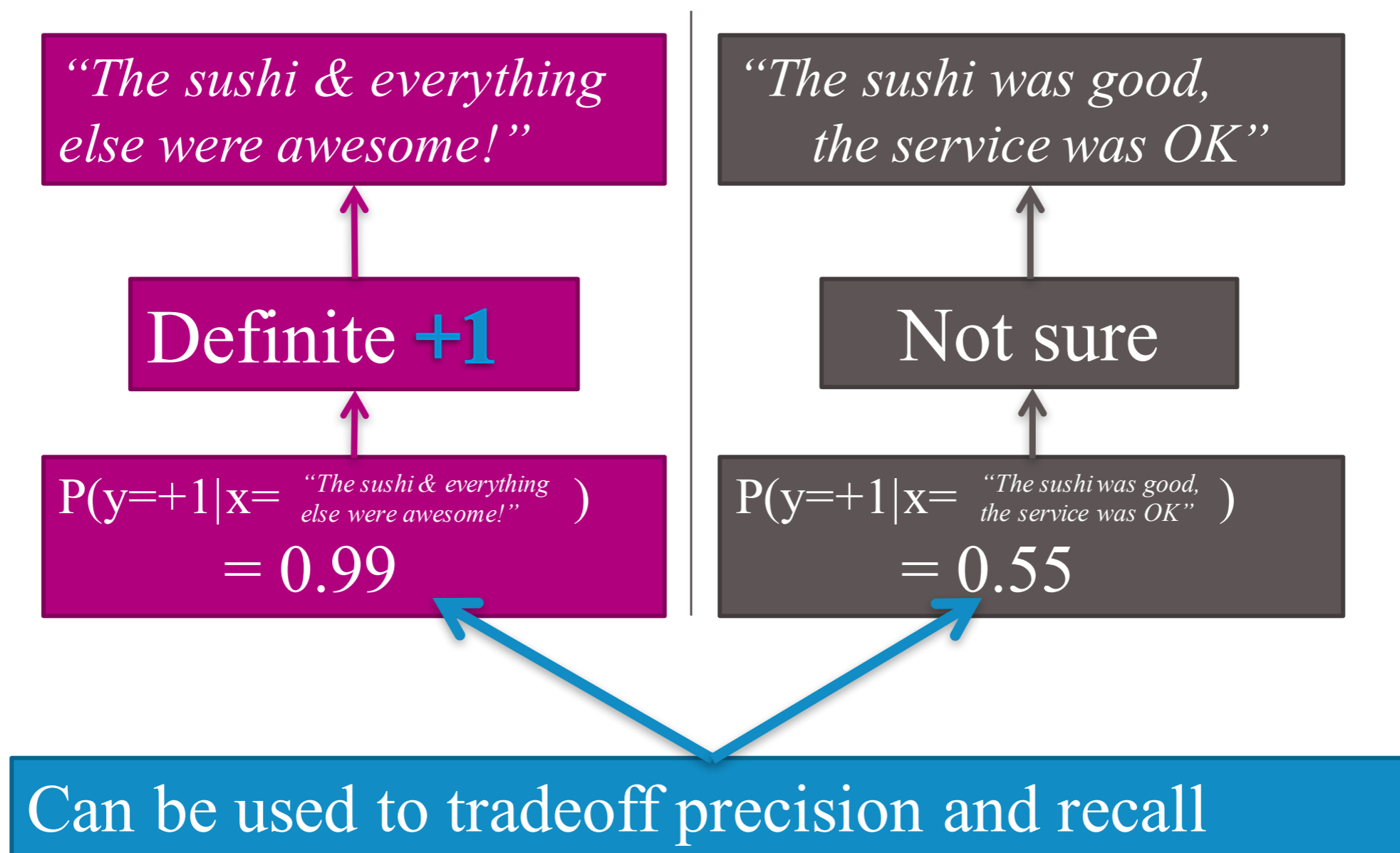| Easily best sushi in Seattle. | 👍 |
The seaweed salad was just OK, vegetable salad was just ordinary.	
I like the interior decoration and the blackboard menu on the wall.	👍
The service is somewhat hectic.	
The sushi was amazing, and the rice is just outstanding.	👍
All the sushi was delicious.	👍
The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

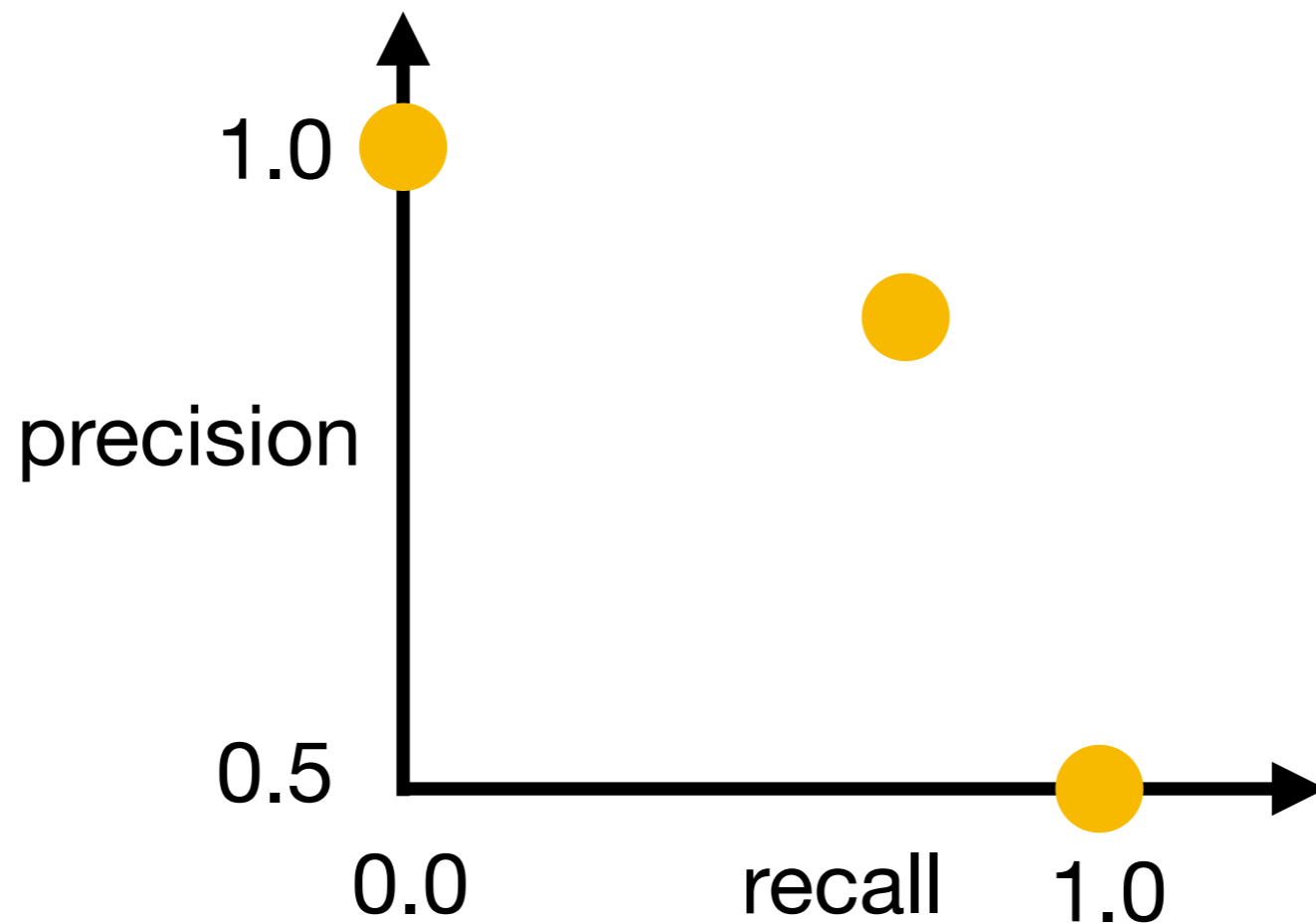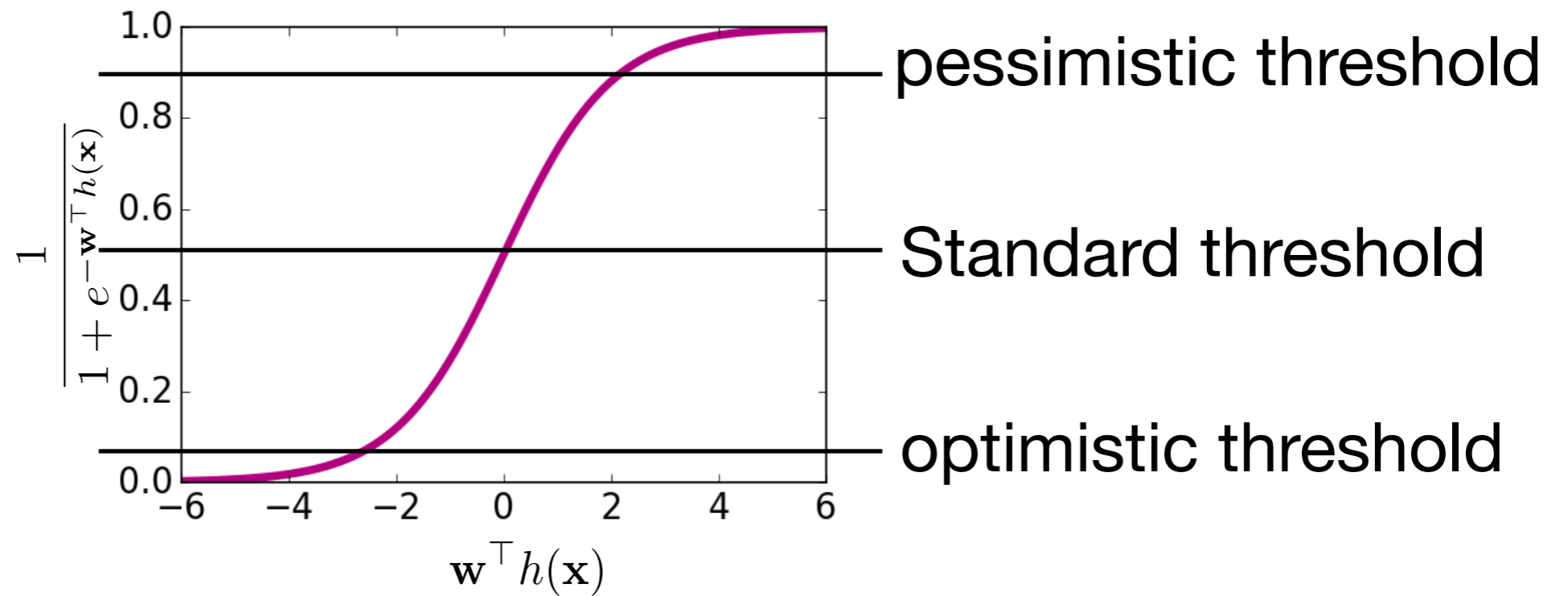| The service is somewhat hectic. | |

10
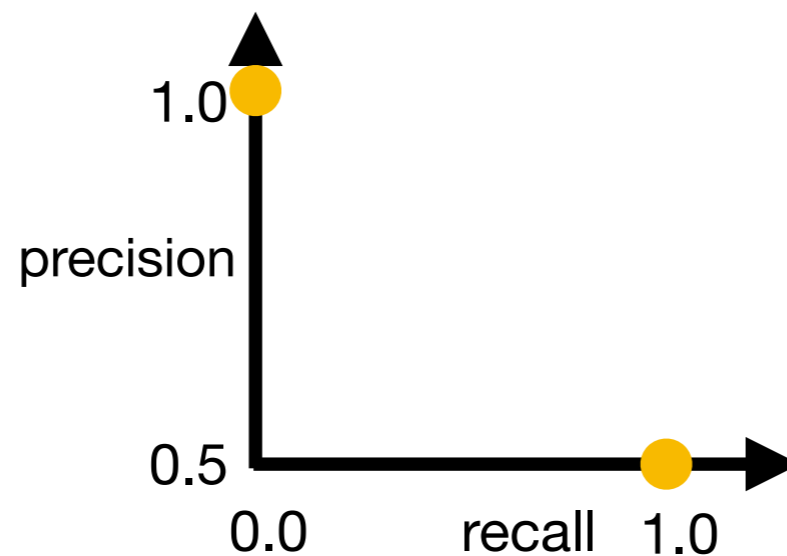
How do we trade-off precision and recall?

- In the case of logistic regression,
  we can choose how many positive predictions we want,
  and then include the higher confident ones first

"The sushi & everything else were awesome!"

"The sushi was good, the service was OK"

Definite **+1**

Not sure

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

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

Can be used to tradeoff precision and recall

logistic regression trade-off

pessimistic threshold

Standard threshold

optimistic threshold

$$\frac{1}{1 + e^{-\mathbf{w}^\top h(\mathbf{x})}}$$

$\mathbf{w}^\top h(\mathbf{x})$

precision

1.0

0.5

0.0   recall   1.0
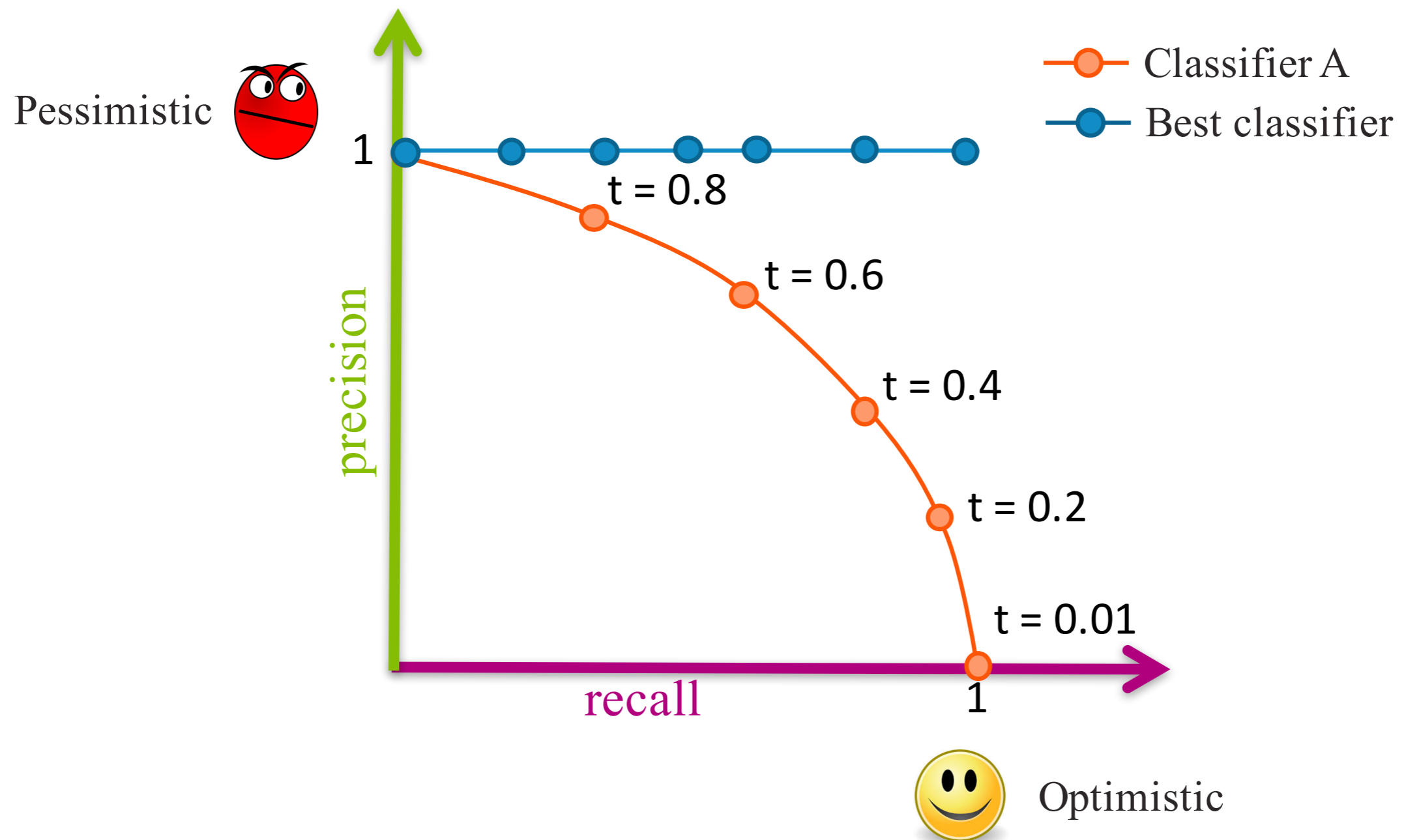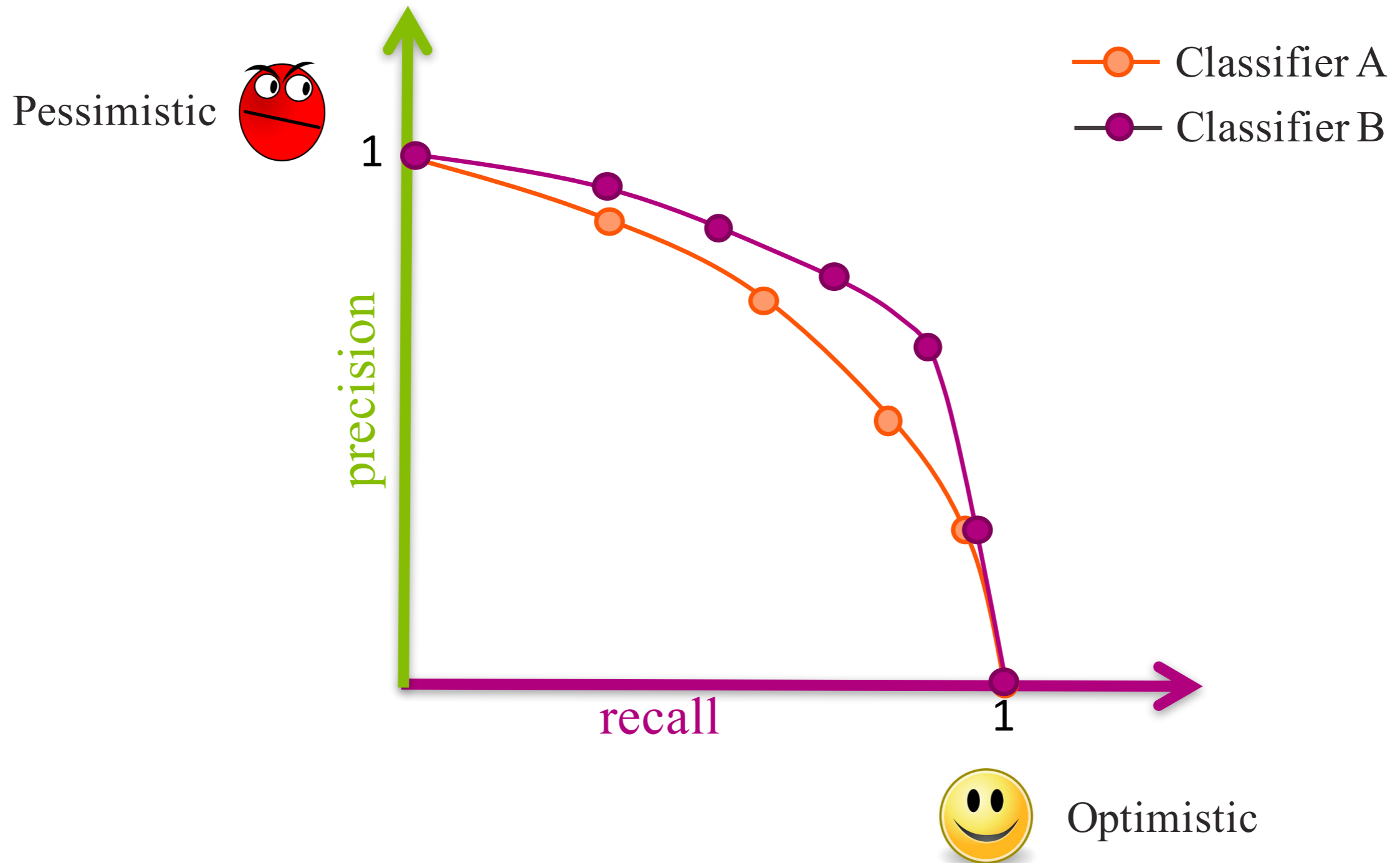
precision-recall curve of a classifier

- given a classifier, we can traverse the thresholds to get a full curve

Pessimistic

Classifier A
Best classifier

1

t = 0.8

t = 0.6

precision

t = 0.4

t = 0.2

t = 0.01

recall   1

Optimistic

- Some times a classifier B is strictly better than classifier A

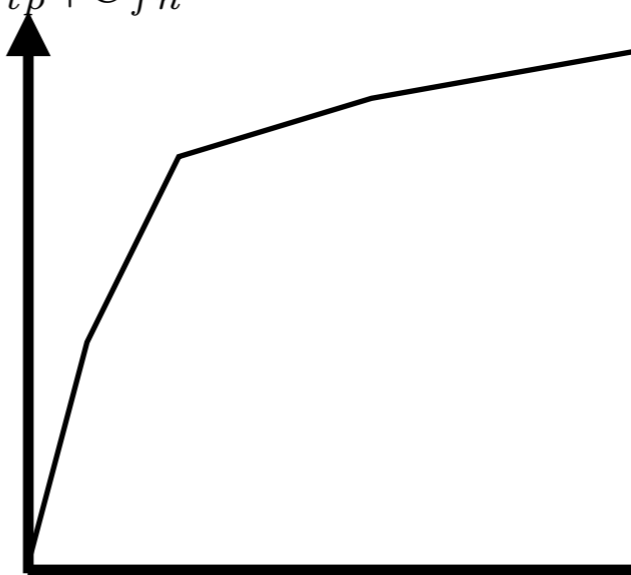- but most of the time, two curves are not directly comparable
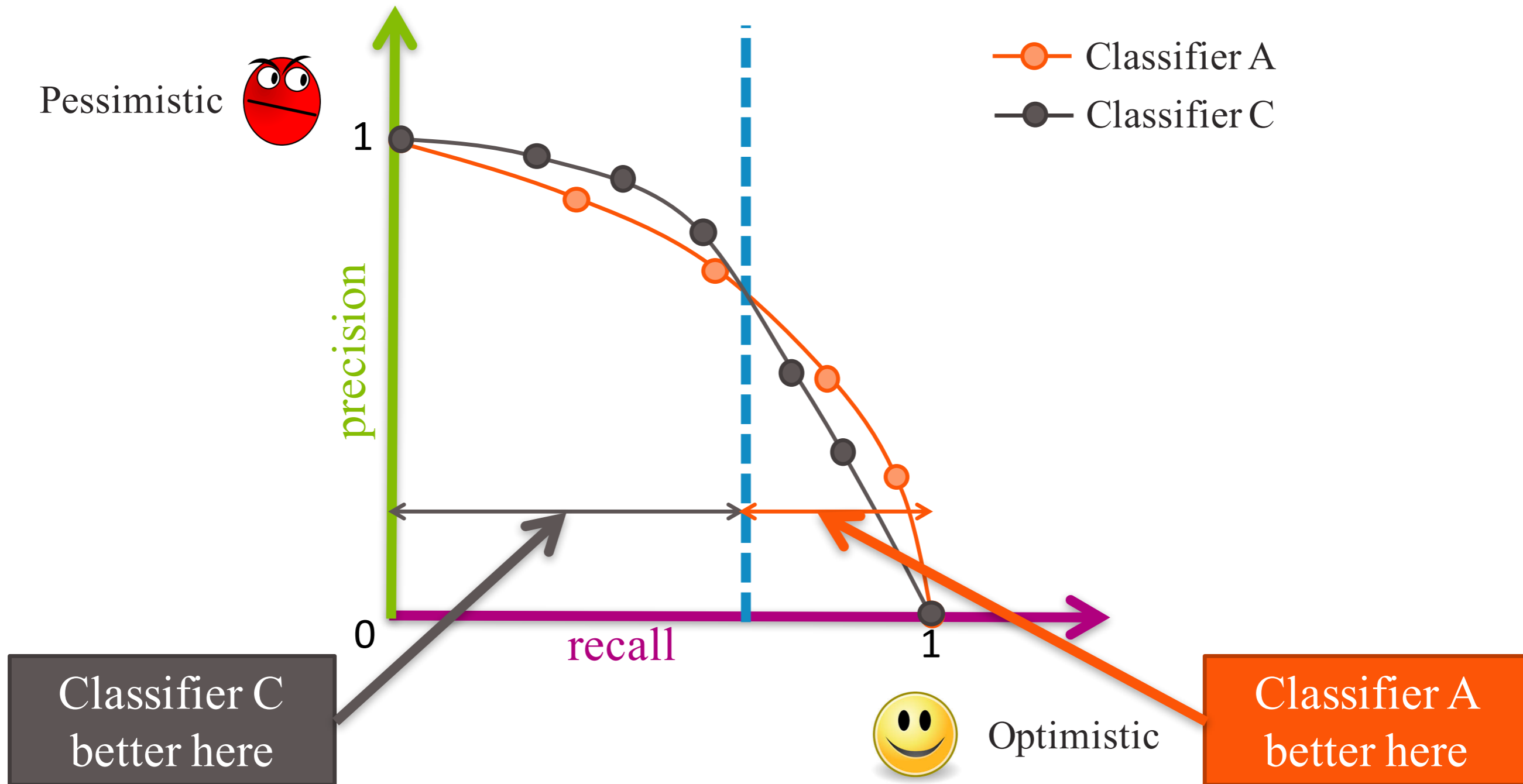
- So for comparisons, people use a single number
  - F1-score $$F1score = 2 \times \frac{precision \times recall}{precision + recall}$$
  - AUC

$$TPR = \frac{C_{tp}}{C_{tp} + C_{fn}} = Recall$$

$$FPR = \frac{C_{fp}}{C_{fp} + C_{tn}}$$

- Or use precision at **k**,
  - precision when top **k** examples are chosen to be predicted as positive

Sentences model
most sure are positive

Showing
k=5 sentences
on website

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