Stanford University

CS 329P : Practical Machine Learning (2021 Fall)

4.1 Evaluation Metrics

Qingqing Huang, Mu Li, Alex Smola

https://c.d2l.ai/stanford-cs329p

Model Metrics

- Loss measures how good the model in predicting the outcome in supervised learning
- Other metrics to evaluate the model performance
  - Model specific: e.g. accuracy for classification, mAP for object detection
  - Business specific: e.g. revenue, inference latency
- We select models by multiple metrics
  - Just like how you choose cars

Stanford CS 329P (2021 Fall) - https://c.d2l.ai/stanford-cs329p

Metrics for Binary Classification

- Accuracy: # correct predictions / # examples

```
sum(y == y_hat) / y.size
```

- Precision: # True positive / # (True positive + False positive)

```
sum((y_hat == 1) & (y == 1)) / sum(y_hat == 1)
```

- Recall: # True positive / # Positive examples

```
sum((y_hat == 1) & (y == 1)) / sum(y == 1)
```

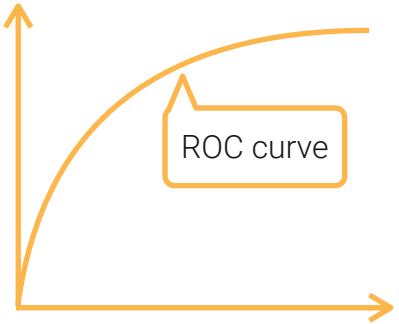
- Be careful of division by 0
- One metric that balances precision and recall
  - F1: the harmonic mean of precision and recall: $2pr/(p+r)$

AUC-ROC

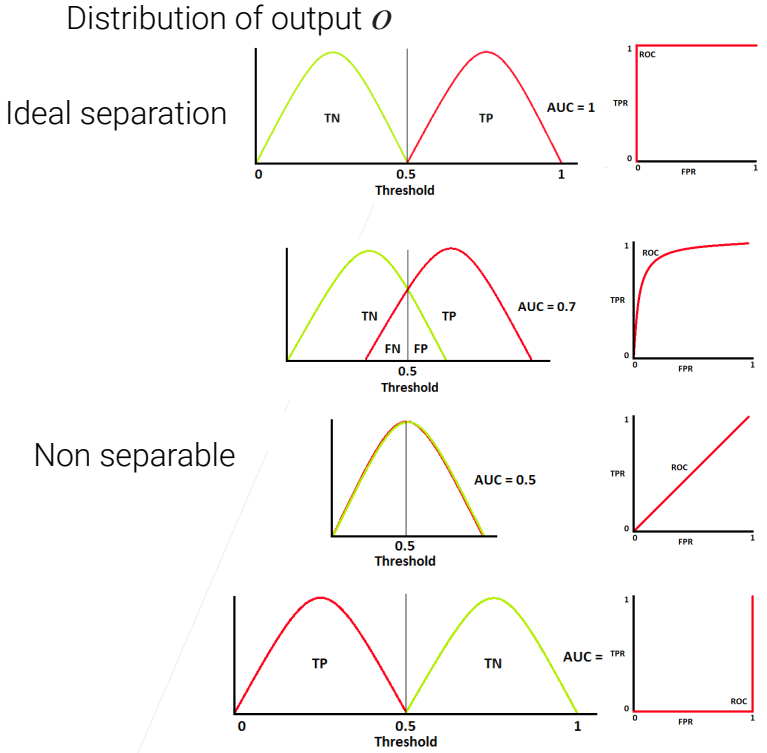
- Measures how well the model can separate the two classes
- Choose decision threshold θ, predict positive if $o \geq \theta$ else neg
- In the range [0.5, 1]

$$\text{TPR} = \frac{\text{\# True Pos predictions}}{\text{\# pos examples}}$$

$$\text{FPR} = \frac{\text{\# False Pos predictions}}{\text{\# neg examples}}$$

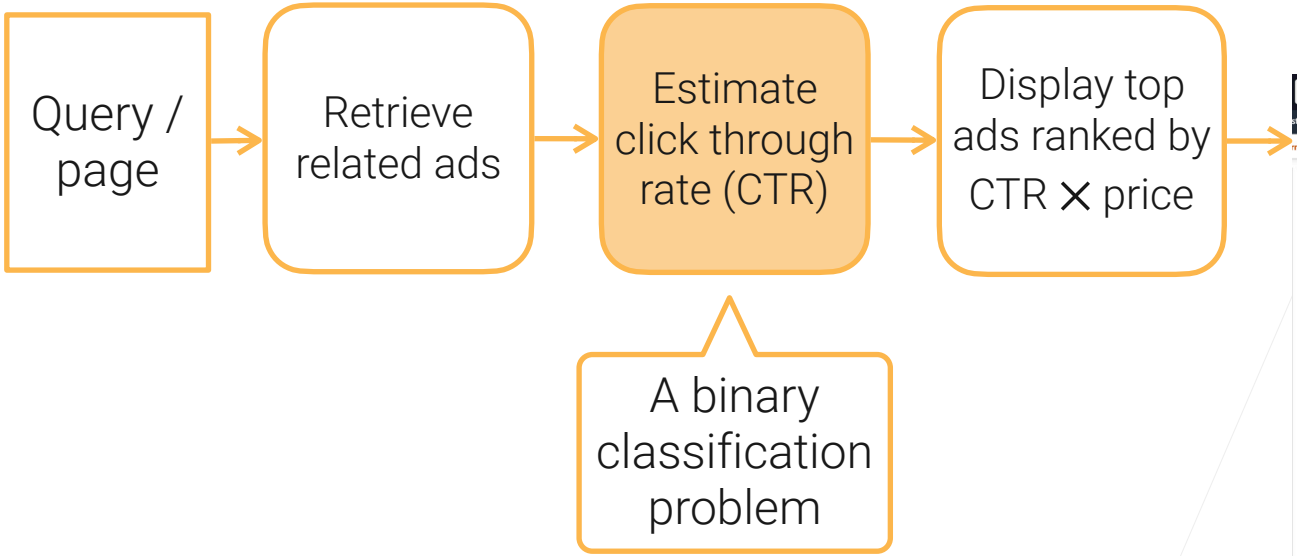


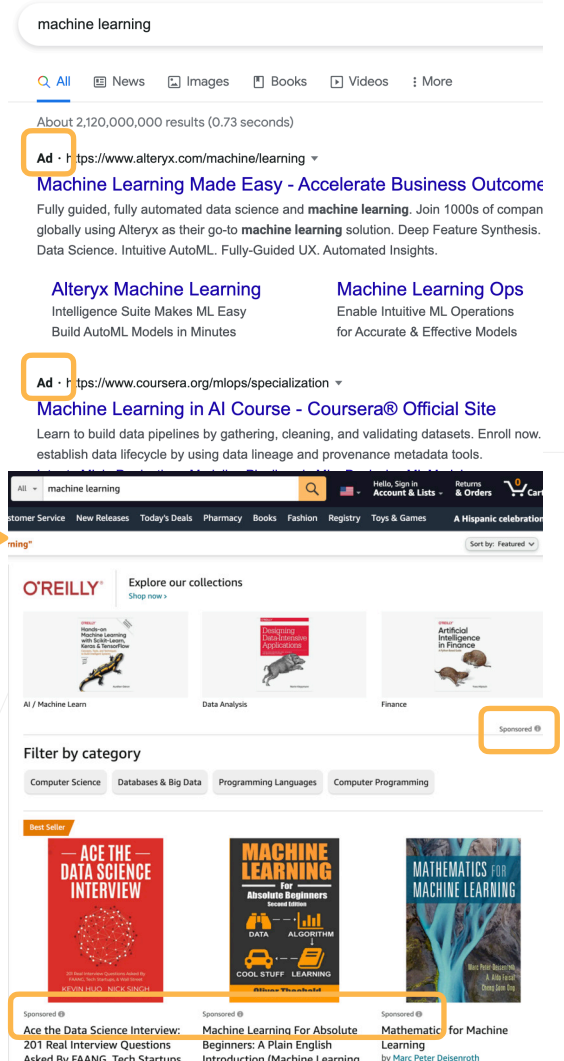
Distribution of output o



Case Study: Displaying Ads

- Ads is one major revenue source for Internet companies





Business Metrics for Displaying Ads

- Optimize both revenue and customer experience
  - Latency: ads should be shown to users at the same time as others
  - ASN: average #ads shown in a page
  - CTR: actual user click through rate
  - ACP: average price advertiser pays per click
- revenue = #pageviews × ASN × CTR × ACP

Displaying Ads: Model → Business Metrics

- The key model metric is AUC
- A new model with increased AUC may harm business metrics, possible reasons:
  - Lower estimated CTR → less ads displayed
  - Lower real CTR because we trained and evaluated on past data
  - Lower prices
- Online experiment: deploy models to evaluate on real traffic data

Summary

- We evaluate models with multiple metrics
- Model metrics evaluate model performance on examples
  - E.g. accuracy, precision, recall, F1, AUC for classification models
- Business metrics measure how models impact the product