Deploying Machine Learning in Production

Alice Zheng and Shawn Scully, Dato
Strata + Hadoop World, London
May 2015
Evaluating Deployed Machine Learning…
What could go wrong?

Alice Zheng and Shawn Scully, Dato
Strata + Hadoop World, London
May 2015
Self introduction

• Background
  – Machine learning research
• Now
  – Build ML tools
  – Teach folks how to use them

@RainyData, @DatoInc
What is Dato?

• A startup based in Seattle, Washington
• Formerly named GraphLab
• We built an ML platform for building and deploying apps
  – Data engineering, ML modeling, deployment to production
  – Graphs, tables, text, images
  – Out of core processing for fast ML on large data
Demo: stratanow.dato.com
What’s in an ML app?

• An application that uses machine learning to make predictions
Machine learning deployment pipeline

Historical data → Prototype model → Deployed model → Predictions

New request → Online adaptive model
Machine learning evaluation

Historical data

Prototype model

Offline evaluation

Deployed model

Online evaluation

New request

Predictions

Online adaptive model

Deployed model

Offline evaluation

Prototype model

Historical data

New request
When/how to evaluate ML

• Offline evaluation
  – Evaluate on historical labeled data

• Online evaluation
  – A/B testing – split off a portion of incoming requests (B) to evaluate new deployment, use the rest as control group (A)
Evaluating ML—What Could Go Wrong?
Evaluation metrics

• Classification
  – Accuracy, precision-recall, AUC, log-loss, etc.

• Ranking
  – Precision-recall, DCG/NDCG, etc.

• Regression
  – RMSE, error quantiles, max error, etc.

• Online models
  – Online loss (error of current model on current example)

Which metric?

- Offline metric ≠ business metric
  - Business metric: customer lifetime value
    - How long does the customer stay on your site?
    - How much more do you sell?
  - Which offline metric does it correspond to?
- Say you are building a recommender
  - “How well can I predict ratings?”
  - Customer sees the first few recommended items
  - Ranking metric is better than rating regression
- Track both business and ML metrics to see if they correlate
Watch out for imbalanced datasets!

Accuracy

100%

99% accuracy!

98% baseline. Doh!

True pos rate

100%

You have 5 positive examples out of 10K. So the important region is only this big. Doh!

false pos rate 100%

Nice ROC curve!
Watch out for rare classes!

When averaging statistics from multiple sources, watch out for different confidence intervals.

Alice
Bob
Catherine
...
Zoe

Average = 0.12

Statistically problematic average of quantities of varying confidence

Correct recommendations
All relevant items
Confidence interval of estimate

A/B testing: T-tests

• Statistical hypothesis testing
  – Is population 1 significantly different from population 0?
• T-tests: are the means of the two populations equal?
• Procedure:
  – Pick significance level \( \alpha \)
  – Compute test statistic
  – Compute p-value (probability of test statistic under the null hypothesis)
  – Reject the null hypothesis if p-value is less than \( \alpha \)
A/B testing: T-tests

• Student’s t-test assumes variances are equal
A/B testing: T-tests

- Welch’s t-test *doesn’t* assume variances are equal
A/B testing: How long to run the test?

• Run the test until you see a significant difference?
  – Wrong! Don’t do this.

• Statistical tests directly control for false positive rate (significance)
  – With probability $1-\alpha$, Population 1 is different from Population 0

• The statistical power of a test controls for the false negative rate
  – How many observations do I need to discern a difference of $\delta$ between the means with power 0.8 and significance 0.05?

• Determine how many observations you need before you start the test
  – Pick the power $\beta$, significance $\alpha$, and magnitude of difference $\delta$
  – Calculate $n$, the number of observations needed
  – Don’t stop the test until you’ve made this many observations
A/B testing: The conundrum of multiple hypotheses

• You are testing 20 models at the same time …
  – … each of them has a 5% chance of being a fluke
  – … on average, expect at least one fluke in this suite of tests

• Adjust the acceptance level when testing multiple hypotheses
  – Bonferroni correction for false discovery rates
A/B testing: Separation of experiences

• How well did you split off group B?

---

Group A

Group B

Unclean separation of experiences!
A/B testing: The shock of newness

- People hate change
  - Why is my button now blue??
- Wait until the “shock of newness” wears off, then measure
- Some population of users are forever wedded to old ways
  - Consider obtaining a fresh population

Click-through rate

The shock of newness

Distribution drift

• Trends and user taste changes over time
  – “I liked house music 10 years ago. Now I like jazz.”

• Models become out of date
  – When to update the model?

• Do both online and offline evaluation
  – Monitor correlation
  – Also useful for tracking business metrics vs. evaluation metrics
Conclusions

• Machine learning are useful in making smart apps
• Evaluating ML models in production is tricky
• Summary of tips:
  – Pick the right metrics
  – Monitor offline and online behavior, track their correlation
  – Be really careful with A/B testing

@RainyData, @DatoInc