UBC MLRG (Summer2017): Online, Active, and Causal Learning

Machine Learning Reading Group (MLRG)

- Machine learning reading group (MLRG) format:
  - Each semester we pick a general topic.
  - Each week someone leads us through a tutorial-style lecture/discussion.
  - So it’s organized a bit more like a “topics course” than reading group.

- We use this format because ML has become a huge field.

Machine Learning Reading Group (MLRG)

- I’ve tried to pack as much as possible into the two ML courses:
  - CPSC 340 covers most of the most-useful methods.
  - CPSC 540 covers most of the background needed to read research papers.

- This reading group covers topics that aren’t yet in these course.
  - Aimed at people who have taken CPSC 340,
    and are comfortable with 540-level material.

Recent MLRG History

- Topics covered in recent tutorial-style MLRG sessions:
  - Summer 2015: Probabilistic graphical models.
  - Fall 2015: Convex optimization.
  - Winter 2016: Bayesian statistics.
  - Summer 2016: Miscellaneous.
  - Fall 2016: Deep learning.
  - Winter 2017: Reinforcement learning.
  - Summer 2017: Online, Active, and Causal Learning ("Time and Actions").

Topic 1: Online Learning

- Usual supervised learning setup:
  - Training phase:
    - Build a model ‘w’ based on IID training examples (x_t, y_t).
  - Testing phase:
    - Use the model to make predictions $\hat{y}_t$ on new IID testing examples $\hat{x}_t$.
    - Our “score” is the total difference between predictions $\hat{y}_t$ and true test labels y_t.

- In online learning there is no separate training/testing phase:
  - We receive a sequence of features x_t.
  - You make prediction $\hat{y}_t$ on each example x_t as it arrives.
    - You only get to see y_t after you’ve made prediction $\hat{y}_t$.
  - Our “score” is the total difference between predictions $\hat{y}_t$ and true labels y_t.
    - We need to predict well as we go (not just at the end).
    - You pay a penalty for having a bad model as you are learning.

Topic 1: Online Learning

- In online learning, we typically don't assume data is IID.
  - Often analyze a weaker notion of performance called "regret".

- Main applications: online ads and spam filtering.

- A common variation is with bandit feedback:
  - There may be multiple possible y_t, we only observe loss for action we choose.
    - You only observe whether they clicked on your ad, not which ads they would have clicked on.
  - Here we have an exploration vs. exploitation trade-off:
    - Should we explore by picking a y_t we don't know much about?
    - Should we exploit by picking a y_t that is likely to be clicked?

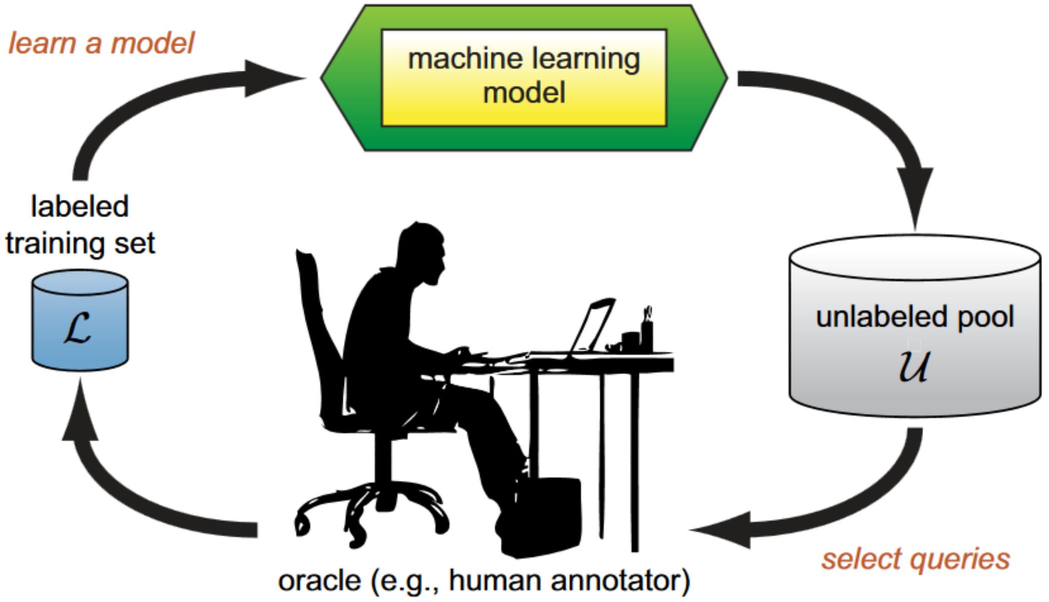
Topic 2: Active Learning

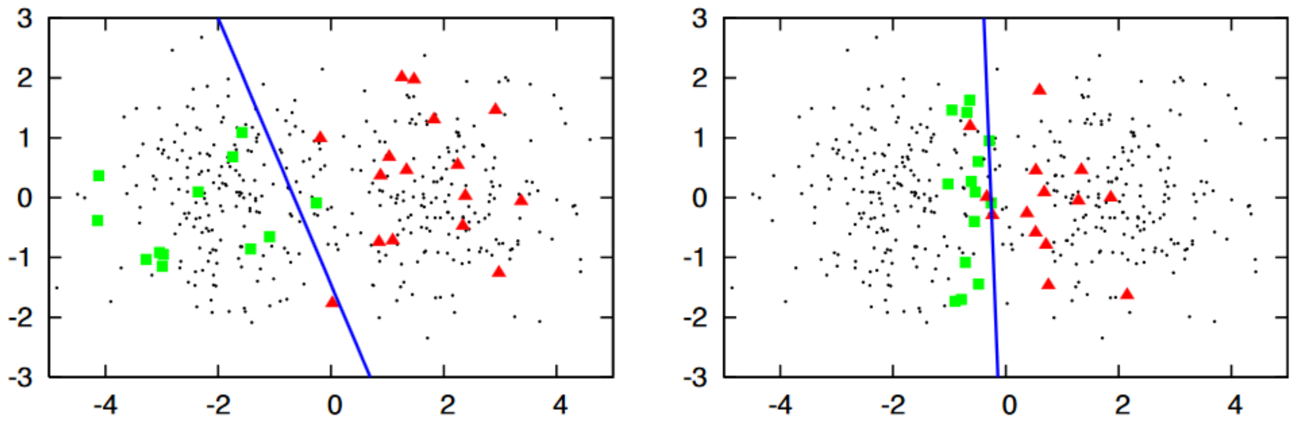
- Supervised learning trains on labeled examples (X,y).
  - The doctor has labeled thousands of images for you.

- Semi-supervised learning trains on (X,y) and unlabeled examples $\tilde{X}$.
  - The doctor has labeled 20 images for you.
  - You have a database of thousands of images.

- Active learning trains only on unlabeled examples $\tilde{X}$.
  - But you can ask the doctor to label 20 images for you.

Topic 2: Active Learning





- Which x^t should we label to learn the most?
- Closely-related to optimal experimental design in statistics.

Topic 3: Causal Learning

- The difference between observational and interventional data:
  - If I see that my watch says 10:55, class is almost over (observational).
  - If I set my watch to say 10:55, it doesn't help (interventional).
- In 340 and 540, we only considered observational data.
  - If our model performs actions, we need to learn effects of actions.
  - Otherwise, it may make stupid predictions.
- We may want to discover direction of causality.
  - "Watch" only predicts of "time" in observational setting (so it's not causal).
  - We can design experiments or make assumptions that find directions.
    - Randomized controledl trials used in medicine.

Topic 3: Causal Learning

- Levels of causal inference:
  - Observational prediction:
    - Do people who take Cold-FX have shorter colds?
  - Causal prediction:
    - Does taking Cold-FX cause you to have shorter colds?
  - Counter-factual prediction:
    - You didn't take Cold-FX and had long cold, would taking it have made it shorter?

- Counter-factuals condition on imaginary pasts.

(pause)

Online Classification with Perceptron

- Perceptron for online linear binary classification [Rosenblatt, 1952]
  - Start with $w_0 = 0$.
  - At time time 't' we receive features x_t.
  - We predict $\hat{y}_t = \text{sign}(w_t^T x_t)$.
  - If $\hat{y}_t \neq y_t$, then set $w_{t+1} = w_t + y_t x_t$.
    - Otherwise, set $w_{t+1} = w_t$.

- Perceptron mistake bound [Novikoff, 1962]:
  - Assume data is linearly-separable with a "margin":
    - There exists w^* with $||w^*||=1$ such that $\text{sign}(x_t^T w^*) = \text{sign}(y_t)$ for all 't' and $|x^T w^*| \geq \gamma$.
  - Then the number of total mistakes is bounded.
    - No requirement that data is IID.

Perceptron Mistake Bound

- Let's normalize each x_t so that $||x_t|| = 1$.
  - Length doesn't change label.
- Whenever we make a mistake, we have $\text{sign}(y_t) \neq \text{sign}(w_t^T x_t)$ and

$$
\begin{aligned}
\|w_{t+1}\|^2 &= \|w_t + y x_t\|^2 \\
&= \|w_t\|^2 + 2\underbrace{y_t w_t^T x_t}_{<0} + 1 \\
&\leq \|w_t\|^2 + 1 \\
&\leq \|w_{t-1}\|^2 + 2 \\
&\leq \|w_{t-2}\|^2 + 3.
\end{aligned}
$$

- So after 'k' errors we have $||w_t||^2 \leq k$.

Perceptron Mistake Bound

- Let's consider a solution w^*, so $\text{sign}(y_t) = \text{sign}(x_t^T w^*)$.
- Whenever we make a mistake, we have:

$$\begin{aligned}
\|w_{t+1}\| &= \|w_{t+1}\|\|w_*\| \\
&\geq w_{t+1}^T w_* \\
&= (w_t + y_t x_t)^T w_* \\
&= w_t^T w_* + y_t x_t^T w_* \\
&= w_t^T w_* + |x_t^T w_*| \\
&\geq w_t^T w_* + \gamma.
\end{aligned}$$

- So after 'k' mistakes we have $||w_t|| \geq \gamma k$.

Perceptron Mistake Bound

- So our two bounds are $||w_t|| \leq \text{sqrt}(k)$ and $||w_t|| \geq \gamma k$.
- This gives $\gamma k \leq \text{sqrt}(k)$, or a maximum of $1/\gamma^2$ mistakes.
- Note that γ is upper-bounded by one due to $||x|| \leq 1$.

Beyond Separable Problems: Follow the Leader

- Perceptron can find perfect classifier for separable data.

- What should we do for non-separable data?
  - And assuming we're not using kernels…

- An obvious strategy is called follow the leader (FTL):
  - At time 't', find the best model from the previous (t-1) examples.
  - Use this model to predict y_t.

- Problems:
  - It might be expensive to find the best model.
    - NP-hard to find best linear classifier for non-separable.
  - It can perform very poorly.

Follow the Leader Counter-Example

- Consider this online convex optimization scenario:
  - At iteration ‘t’, we make a prediction w_t.
  - We then receive a convex function f_t and pay the penalty $f_t(w_t)$.
    - f_t could be the logistic loss on example ‘t’.

- In this setting, follow the leader (FTL) would choose:

$$w_t \in \operatorname{argmin}_w \sum_{i=1}^{t-1} f_i(w).$$

- The problem is convex but the performance can be arbitrarily bad…

Follow the Leader Counter Example

- Assume $x \in [-1,1]$ and:
  - $f_1(x_1) = (1/2)x^2$.
  - $f_2(x_2) = -x$.
  - $f_3(x_3) = x$.
  - $f_4(x_4) = -x$.
  - $f_5(x_5) = x$.
  - $f_6(x_6) = -x$.
  - $f_7(x_7) = x$.
  - …
- FTL objective:
  - $F_1(x_1) = (1/2)x^2$.
  - $F_2(x_2) = -(1/2)x^2$.
  - $F_3(x_3) = (1/2)x^2$.
  - $F_4(x_4) = -(1/2)x^2$.
  - $F_5(x_5) = (1/2)x^2$.
  - $F_6(x_6) = -(1/2)x^2$.
  - $F_7(x_7) = (1/2)x^2$.
  - …
- FTL predictions:
  - x_1 = (initial guess)
  - $x_2 = 0$
  - $x_3 = 1$ (worst possible)
  - $x_4 = -1$ (worst possible)
  - $x_5 = 1$ (worst possible)
  - $x_6 = -1$ (worst possible)
  - $x_7 = 1$ (worst possible)
  - …

Regularized FTL and Regret

- Worst possible sequence:
  - {+1,-1,+1,-1,+1,-1,+1,-1,…}
- FTL produces the sequence:
  - {x0,0,+1,-1,+1,-1,+1,-1,…}, which is close to the worst possible.
- Best possible sequence:
  - {0,+1,-1,+1,-1,+1,-,1,+1,…}
- Best sequence with a fixed prediction:
  - {0,0,0,0,0,0,0,0,…}

- We have no way to bound error compared to best sequence: could have adversary.

- We instead consider a weaker notion of "success" called regret:
  - How much worse is our total error than optimal fixed prediction at time 't'.
  - Note that fixed prediction might change with 't'.

- Next week we'll see algorithms with optimal regret.

Schedule

Date	Topic	Presenter
Jun 6	Motivation/overview, perceptron, follow the leader.	Mark
Jun 13	Online convex optimization, mirror descent	Julie
Jun 20	Multi-armed bandits, contextual bandits	Alireza
Jun 27	Heavy hitters	Michael
Jul 4	Regularized FTL, AdaGrad, Adam, online-to-batch	Raunak
Jul 11	Best-arm identification, dueling bandits	Glen
Jul 18	Uncertainty sampling, variance/error reduction, QBC	Nasim
Jul 25	A/B testing, Optimal experimental design	Mohamed
Aug 1	Randomized controlled trials, do-calculus	Sanna
Aug 8	Granger causality, independent component analysis	Issam
Aug 15	Counterfactuals	Eric
Aug 22	MPI causality	Julieta
Aug 29	Instrumental variables	Jimmy