CIS 520 Machine Learning Summary

What we covered
What we didn’t cover
interpretation, causality

Lyle Ungar

Review Session
Final
Course goals

◆ Be familiar with all major ML methods
  ● Regression (linear, logistic), regularization, feature selection
  ● K-NN, Decision trees, Random Forests, SVMs
  ● PCA, K-means, GMM, Autoencoders
  ● Naive Bayes, Bayes Nets, Markov Nets, LDA, HMMs
  ● Boosting, perceptrons, LMS
  ● Deep learning (CNNs, GANs)
  ● Reinforcement Learning (MDP, Q-learning)

◆ Know their strengths and weaknesses
  ● know jargon, concepts, theory
  ● be able to modify and code algorithms
  ● be able to read current literature

We did all of these except one!
Components of ML

- **Representation**
  - Feature set
  - Model form

- **Loss function**
  - And regularization penalty

- **Optimization method**
  - For parameter estimation
  - For model selection and hyperparameter tuning
Representations

- **Linear models**
  - Hyperplane as a separator
  - Kernel methods
- **Decision Trees**
  - Random forests, gradient tree boosting
- **Neural nets**
  - CNNs, GANs and Conditional GANs
  - Recurrent Nets/LSTMs
- **Structured X, y: Graphs & Trees (not covered)**
Loss Functions

- $L_0$
- Hinge
- Logistic
- Exponential (adaboost)
Regularization priors

\[
\text{Argmin}_w \| y - w \cdot x \|_2^2 + \lambda \| w \|_p^p
\]

- **L_2** \|w\|_2^2
  - Gaussian prior: \( p(w) \sim \exp(-|w|^2/\sigma^2) \)
- **L_1** \|w\|_1
  - Laplace prior: roughly \( p(w) \sim \exp(-|w|/\sigma^2) \)
- **L_0** \|w\|_0
  - Spike and slab prior

\[
\log P(D_x, D_y, \theta) = \log P(D_x, D_y | \theta) + \log P(\theta) = -\text{loss}(\theta) + \text{regularizer}(\theta)
\]
If the x’s have been standardized (mean zero, variance 1) then we can visualize the shrinkage:

- $L_2$ = Ridge
  - sum of squares
  - Shrunk $w$

- $L_1$ = Lasso
  - sum of abs value
  - Shrunk $w$

- $L_0$ = “stepwise regression”
  - Number of features
  - Shrunk $w$
Bias-Variance Trade-off

\[ E_{x,y,D}[(h(x; D) - y)^2] = \]

\[ E_{x,D}[(h(x; D) - \bar{h}(x))^2] + E_x[(\bar{h}(x) - \bar{y}(x))^2] + E_{x,y}[\overline{(y(x) - y)^2}] \]

- Variance
- Bias^2
- Noise
Optimization methods

- **Gradient descent**: Stochastic, minibatch
- **Closed form** (e.g. \( w = (X^T X)^{-1} X^T y \))
- **Search**: streamwise, stepwise, stagewise
- **Power method** (for eigenvectors)
- **Lagrange Multipliers** (constrained optimization)
  - not really covered
Optimization methods

◆ EM
  - E: expected value of hidden values
  - M: MLE or MAP estimate of parameters

◆ Other alternating methods
  - X \sim SW^T \text{ for ICA, NNMF (non-negative matrix factorization)}
Hyperparameter Optimization

- **Search**
  - e.g., $L_1$, $L_2$, elastic net hyperparameters
  - Neural network structure, regularization

- **Auto-SKlearn**
  - Initialize hyperparameters from model predicting accuracy as a function of problem description and hyperparameter values

- **Auto-ML**
  - Use reinforcement learning to learn a ‘design policy’
Distance and similarity

- **Distances induced from norms**
  - \( \|x_1 - x_2\|_0 \), \( \|x_1 - x_2\|_1 \), \( \|x_1 - x_2\|_2 \), ...

- **Similarities from kernels**
  - \( k(x_1, x_2) \)

- **Probability-based Divergences**
  - \( D_{KL}(p||q) = \sum_k p_k \log(p_k/q_k) \) - KL-divergence
  - \( H(p, q) = H(p) + D_{KL}(p||q) \) - cross-entropy
    
    \[ = - \sum_k p_k \log(q_k) \]
    
    - \( p \) is the true distribution, \( q \) is the approximate
Cross entropy and log-likelihood

- **Cross-entropy**
  - $H(p, q) = - \sum_k p_k \log(q_k)$ summed over labels $k$
  - $- \sum_i \sum_k \delta_{ik} \log(p(y_i = k | x = x_i))$ \hspace{1cm} $\delta_{ik} = 1$ iff $y_i = k$
  - $-\text{Sum of the estimated log probabilities of the true answers}$
  - $\log \prod_i p(y_i | x_i) = \sum_i \log p(y_i | x_i)$ log-likelihood
KL-Divergence

\[ D_{KL}(p||q) = \sum_k p_k \log(p_k/q_k) \]

\textbf{Mutual information}

\[ \text{MI}(X,Y) = D_{KL}(P(X,Y) \ || \ P(X)Q(X)) \]

\textbf{Information gain}

\[ IG(Y|X_j) = D_{KL}(P(Y|X_j) \ || \ P(Y)) = H(Y) - H(Y|X_j) \]

Which feature \( X_j \) will maximize the information gain?

\textbf{Bayesian Experimental Design}

For which \( x \) will the label \( y \) (in expectation) most change \( p(w) \)

https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence
Types of Learning

- **Supervised** \( X, y \)
  - Given an observation \( x \), what is the best label \( y \)?

- **Unsupervised** \( X \)
  - Given a set of \( x \)'s, cluster or summarize them

- **Reinforcement**
  - Given a sequence of states \( x \) and possible actions \( a \), learn which actions maximize reward.

What kind of learning is missing here?
Unsupervised methods

- **PCA, ICA, NNMF**
  - $X \sim S V^T$

- **K-means, GMM**

- **Auto-encoders**
  - Information bottleneck
  - Denoising
  - Variational (not covered)

Most of these minimize reconstruction error subject to some constraints
Reinforcement learning

- **Model-based**
  - MDP

- **Model-free**
  - **Shallow**: TD(0) vs. **Deep**: Monte-Carlo
  - **Value**: $V(s)$ vs. **Q-learning**: $Q(s,a)$

- **On policy** ($\varepsilon$-greedy) vs. **off-policy**
  - Trade off *exploration* and *exploitation*
Model Interpretation

◆ Global: What does this model do?
  ● Variable importance
    ■ E.g. how much does accuracy suffer if you remove it?
◆ Local: Why did you make this decision for this x?
  ● Decision tree: path taken
  ● LIME (Local Interpretable Model-Agnostic Explanations)
LIME algorithm

- Sample instances near the target instance
- Predict labels using full model
- Fit a sparse locally weighted regression
- The dashed line is the “explanation”
LIME

- Do local perturbations to $x$

Original Image    Interpretable Components

- Fit locally weighted model

Why Should I Trust You?: Explaining the Predictions of Any Classifier
Ribeiro Singh & Guestrin

LIME

Generate a data set of perturbed instances by turning some of the interpretable components “off” (gray)

**Perturbed Instances** | **P(tree frog)**
--- | ---
![Perturbed Image 1] | 0.85
![Perturbed Image 2] | 0.00001
![Perturbed Image 3] | 0.52

Original Image
P(tree frog) = 0.54

Locally weighted regression

Explanation = pixels with high weights
LIME explains alternate predictions

LIME

- Works on SVMs, Random forests, Nnets..
- Works on text or images

Prediction probabilities

<table>
<thead>
<tr>
<th></th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>atheism</td>
<td>0.58</td>
</tr>
<tr>
<td>christian</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11

NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.

Correlation and Causality

I used to think correlation implied causation.

Then I took a statistics class. Now I don’t.

Sounds like the class helped. Well, maybe.

https://xkcd.com/552/
High correlation between...

- Radio ownership and population in insane asylums
  - England, 20th century
- Daily ice cream consumption and rape incidents
  - US, 21st century
- Stork population and babies born
  - Germany, 20th century
Storks and Babies

New evidence for the theory of the stork.

- Höfer T, Przyrembel H, Verleger S.

Data from Berlin (Germany) show a significant correlation between the increase in the stork population around the city and the increase in [baby] deliveries outside city hospitals (out-of-hospital deliveries). However, there is no correlation between deliveries in hospital buildings (clinical deliveries) and the stork population. The decline in the number of pairs of storks in the German state of Lower Saxony between 1970 and 1985 correlated with the decrease of deliveries in that area.
Feedback complicates causality

- Room temperature as a function of whether the heat is on.
Causality Matters

- If treatment is different based on gender
  - Probability of recovery based on seeing that someone is treated is different than the probability if one knows the gender.

- Causality is usually impossible to infer
  - Does treatment cause high blood pressure or high blood pressure cause treatment?
Questions

- My decision tree indicates that zip code, house price, having aluminum siding, and owning a boat are all useful predictors of purchase of projection TVs, while income is not.
  - What might be going on?
  - Is this a problem?

- Predicted polymer quality does not depend on the temperature of the reactor
  - What might be going on?
  - Is this a problem?
Questions

- Rich Caruana found that among patients with pneumonia, those with asthma had a lower chance of dying.
  - What might be going on?
  - Is this a problem?
What to use when? - Supervised

- $n=100, \ p=100,000$  
  - PCR, Ridge, SVM semi-supervised

- $n=10,000, \ p=10,000$  
  - Anything in sklearn

- $n=100,000, \ p=100$  
  - Sklearn, Nnet

Ask: What structure do I expect to see?
What to use when? - Unsupervised

- $n=100, \ p=100,000 \quad K$-means or PCA
- $n=10,000, \ p=10,000 \quad K$-means or PCA
- $n=100,000, \ p=100 \quad K$-means? GMM? Autoencoder?
The biggest open problem: Transfer learning

- Embeddings
  - Images, words
- Multitask learning
- One shot learning
- Learn “deep structure”
  - Grammar
  - Causal models
  - ???
See many of you for the review session
See all of you for the final
Stay in touch & let me know how you use ML …

Thank you!!!