Active Learning & Experimental Design

Daniel Ting

Heavily modified, of course, by Lyle Ungar

Original Slides by Barbara Engelhardt and Alex Shyr

Lyle Ungar, University of Pennsylvania
Motivation

- **Data collection may be expensive**
  - Cost of time and materials for an experiment
  - Cheap vs. expensive data
    - Raw images vs. annotated images

- **Not all labels are equally useful**

- **Want to collect the “best” data at minimal cost**

Where do labels come from? What observations should one label?
Toy examples

Assume you are learning $y = f(x)$ for $x$ a scalar

You are learning linear regression on $[-1,1]$. You can pick two $x$’s to get $y$’s for.

What two values would you pick?
A) $-1/3, 1/3$
B) $-1, 1$
C) $0, 1$
D) Something else
Toy examples

Assume you are learning $y = f(x)$ for $x$ a scalar
You are learning an SVM classifier on [-1.1].
You can pick 4 $x$’s to get $y$’s for.

What strategy would you use to pick $x$’s?

A) Pick -1, -1/3, 1/3, 1
B) Pick -1, 1, see what the answer is, then pick next $x$
C) Pick -1/3, 1/3, see what the answer is, then pick next $x$
D) Something else
Toy Example: 1D classifier

Unlabeled data: labels are all 0 then all 1 (left to right)
Classifier (threshold function): \( h_w(x) = 1 \) if \( x > w \) (0 otherwise)
Goal: find transition between 0 and 1 labels in minimum steps
Naïve method: choose points to label at random on line
  • Requires \( O(n) \) training data to find underlying classifier
Better method: binary search for transition between 0 and 1
  • Requires \( O(\log n) \) training data to find underlying classifier
  • Exponential reduction in training data size!
Example: collaborative filtering

- Users usually rate only a few movies
  - ratings are “expensive”
- Which movies do you show users to best extrapolate movie preferences?
Example: collaborative filtering

- Users usually rate only a few movies
  - ratings are “expensive”
- Which movies do you show users to best extrapolate movie preferences?

- **Baseline algorithms:**
  - Random: \( m \) movies randomly
  - Most Popular Movies: \( m \) most frequently rated movies
- **Most popular movies is not better than random design!**
- **Popular movies rated highly by all users; do not discriminate tastes**

[Yu et al. 2006]
What criteria might you use to select which items to label?
Active Learning Methods

- Active learning
  - Uncertainty sampling
  - Query by committee
  - Information-based loss functions
- Optimal experimental design
- Non-linear optimal experimental design
- Summary
Active learning

- **Setup:** Given existing knowledge \((X,y)\), choose where to collect more data
  - Assume access to cheap unlabeled points
  - Make a query to obtain expensive label
  - Want to find labels that are “informative”
- **Output:** Classifier / predictor trained on less labeled data
- **Similar to “active learning” in classrooms**
  - Students ask questions, receive a response, and ask further questions
  - Contrast: passive learning: student just listens to lecturer
- **This lecture covers:**
  - how to measure the value of data
  - algorithms to choose the data
Example: Gene expression and Cancer classification

- Active learning takes 31 points to achieve same accuracy as passive learning with 174
Active Learning Setup

- Active learner picks which data point $x$ to query
- Receive response $y$ from an oracle
- Update parameters $q$ of the model
- Repeat

- Query selected to minimize some loss function ("risk")
Active Learning

◆ Heuristic methods for reducing risk:
  ● Select “most uncertain” data point
  ● Select “most informative” data point
    ▪ for distinguishing different hypotheses
    ▪ to optimize expected information gain

◆ Some computational considerations:
  ● May be many queries to calculate loss for
    ▪ Subsample points
    ▪ Probability far from the true min decreases exponentially
  ● May not be easy to calculate loss
Uncertainty Sampling

- Query the item (x) that the current classifier is most uncertain about
- Needs measure of uncertainty
- Examples:
  - Entropy
  - Least confident predicted label
  - Euclidean distance (e.g. point closest to margin in SVM)
Example: Gene expression and Cancer classification

◆ Data: Cancerous lung tissue samples
  ● “Cheap” unlabeled data
    ■ gene expression profiles from Affymetrix microarray
  ● Labeled data:
    ■ 0-1 label for adenocarcinoma or malignant pleural mesothelioma

◆ Method:
  ● Linear SVM
  ● Measure of uncertainty
    ■ distance to SVM hyperplane
Example: Gene expression and Cancer classification

Active learning takes 31 points to achieve same accuracy as passive learning with 174
Query by Committee

◆ Which unlabeled point should you choose?

a) Red arrow point
b) Green arrow point
Query by Committee

- Yellow = valid hypotheses
Query by Committee

- Point on max-margin hyperplane does not reduce the number of valid hypotheses by much
Query by Committee

- Queries an example based on the degree of disagreement between committee of classifiers
Query by Committee

- Prior distribution over classifiers/hypotheses
- Sample a set of classifiers from distribution

- Natural for ensemble methods which are already samples
  - Random forests, Bagged classifiers, etc.

- Measures of disagreement
  - Entropy of predicted responses
  - KL-divergence of predictive distributions
Query by Committee Application

- Use Naïve Bayes model for text classification (20 Newsgroups dataset)

[McCallum & Nigam, 1998]
Information-based Loss Function

◆ Previous methods looked at uncertainty at a single point
  ● Does not look at whether you can actually reduce uncertainty or if adding the point makes a difference in the model

◆ Want to model notions of information gained
  ● Maximize KL divergence between posterior and prior
    \[ KL(P || \pi) = \text{# of bits gained about model} \]
  ● Maximize reduction in model entropy between posterior and prior (reduce number of bits required to describe distribution)

◆ All of these can be extended to optimal design algorithms
◆ Must decide how to handle uncertainty about query response, model parameters

[MacKay, 1992]
Entropy

- A measure of information in random event $X$ with possible outcomes \{${x_1, \ldots, x_n}$\}

$$H(x) = - \sum_{i} p(x_i) \log_2 p(x_i)$$

- The average minimum number of yes/no questions to answer some question
  - Related to binary search

[Shannon, 1948]
Kullback Leibler divergence

- \( P \) = true distribution;
- \( Q \) = alternative distribution that is used to encode data
- KL divergence is the expected extra message length per datum that must be transmitted using \( Q \)

\[
D_{KL}(P \parallel Q) = \sum_i P(x_i) \log \left( \frac{P(x_i)}{Q(x_i)} \right)
\]

\[
= - \sum_i P(x_i) \log Q(x_i) + \sum_i P(x_i) \log P(x_i)
\]

\[
= H(P,Q) - H(P)
\]

= Cross-entropy - entropy

- Measures how different the two distributions are
KL divergence properties

- **Non-negative:** $D(P||Q) \geq 0$

- **Divergence 0 if and only if $P$ and $Q$ are equal:**
  - $D(P||Q) = 0$ iff $P = Q$

- **Non-symmetric:** $D(P||Q) \neq D(Q||P)$

- **Does not satisfy triangle inequality**
  - $D(P||Q) \leq D(P||R) + D(R||Q)$
KL divergence properties

◆ Non-negative: \( D(P||Q) \geq 0 \)

◆ Divergence 0 if and only if \( P \) and \( Q \) are equal:
  - \( D(P||Q) = 0 \) iff \( P = Q \)

◆ Non-symmetric: \( D(P||Q) \neq D(Q||P) \)

◆ Does not satisfy triangle inequality
  - \( D(P||Q) \nless D(P||R) + D(R||Q) \)

Not a distance metric
KL divergence as info gain

- The KL divergence of the posteriors measures the information gain expected from query \((x')\):

\[
D( p(\theta | x, x') \parallel p(\theta | x))
\]

- Goal: choose a query that maximizes the KL divergence between the updated posterior probability and the current posterior probability
  - This represents the largest expected information gain
Active learning warning

- Choice of data is only as good as the model itself
- Assume a linear model, then two data points are sufficient
- What happens when data are not linear?
Active Learning = Sequential Experimental Design

- Active learning
  - Uncertainty sampling
  - Query by committee
  - Information-based loss functions

- Optimal experimental design
  - A-optimal design
  - D-optimal design
  - E-optimal design
  - Non-linear optimal experimental design

- Summary
Experimental Design

- Many considerations in designing an experiment
  - Dealing with confounders
  - Feasibility
  - Choice of variables to measure
  - Size of experiment ( # of data points )
  - Conduction of experiment
  - Choice of interventions/queries to make
  - Etc.
Experimental Design

Many considerations in designing an experiment
- Dealing with confounders
- Feasibility
- Choice of variables to measure
- Size of experiment (# of data points)
- Conduction of experiment
- Choice of interventions/queries to make
- Etc.

We will only look at one of them
Optimal Experimental Design?

- Heuristics give empirically good performance but
  - Not that much theory on how good the heuristics are

- Optimal experimental design gives
  - theoretical criteria for choosing a set of points to label
  - for a specific set of assumptions and objectives

- Theory is good when you only get to run (a series of) experiments once
Optimal Experimental Design

- Given a model $M$ with parameters $\beta$,
  - What queries are maximally informative
    i.e. will yield the best estimate of $\beta$
- "Best" minimizes variance of estimate
  - Equivalently, maximizes the Fisher Information
    \[ I(\hat{\beta}) \approx \text{var}(\hat{\beta})^{-1} \text{ if } \hat{\beta} \text{ is the mle} \]
  - Different methods use different ways of computing a scalar from the variance matrix
- Linear models
  - Optimal design does not depend on $\beta$!
- Non-linear models
  - Depends on $\beta$, but can use Taylor expansion to linear model
Goal: Minimize variance of $w$

If $y = x^T\beta + \varepsilon$ then $w = (X^TX)^{-1}X^Ty$

$w \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1})$

We want to minimize the variance of our parameter estimate $w$, so pick training data $X$ to minimize $(X^TX)^{-1}$

But that is a matrix, so we need to reduce it to a scalar

- **A-optimal** (average) design minimizes $\text{trace}(X^TX)^{-1}$
- **D-optimal** (determinant) design minimizes $\log \det(X^TX)^{-1}$
- **E-optimal** (extreme) design minimizes $\max$ eigenvalue of $(X^TX)^{-1}$

Alphabet soup of other criteria (C-, G-, L-, V-, etc.)
Fisher Information Matrix

The negative of the expected value of the Hessian of the log of the probability of an observation given the parameters $\theta$.

$$(I(\theta))_{i,j} = -E \left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(X; \theta) \mid \theta \right]$$

If $y = x^T w + \epsilon$
Then $p(y_t) \sim \exp(-(x_t^T w)^2/2\sigma^2)$

$m_{ij} = -E[d^2dw_i dw_j (\log \exp(-(x_t^T w)^2/2\sigma^2))]$

$= E[d^2dw_i dw_j ((x_t^T w)^2/2\sigma^2)]$

$\sim E[x_i^T x_j]$
So $M \sim X^TX$
A-Optimal Design

- **A-optimal** design minimizes the trace of $(X^TX)^{-1}$
  - Minimizing trace (sum of diagonal elements) essentially chooses maximally independent columns (small correlations between interventions)
- **Tends to choose points on the border of the dataset**

Example: mixture of four Gaussians

[Yu et al., 2006]
**A-Optimal Design**

- *A-optimal* design minimizes the trace of $(X^TX)^{-1}$

Example: 20 candidate data points, minimal ellipsoid that contains all points
D-Optimal design

- **D-optimal** design minimizes log determinant of $(X^TX)^{-1}$
- Equivalent to
  - choosing the confidence ellipsoid with minimum volume
    (“most powerful” hypothesis test in some sense)
  - Minimizing entropy of the estimated parameters
- Most commonly used optimal design

[Boyd & Vandenberghe, 2004]
E-Optimal design

- *E-optimal* design minimizes largest eigenvalue of \((X^TX)^{-1}\)
- Minimizes the diameter of the confidence ellipsoid
Summary of Optimal Design

Lyle H Ungar, University of Pennsylvania

[Boyd & Vandenberghe, 2004]
Practicalities

- Sometimes you can generate an $x$ arbitrarily
- More often you need to select from a set of given $x$’s
  - This can be an expensive search!
Experimental Design

- Introduction: information theory
- Active learning
  - Uncertainty sampling
  - Query by committee
  - Information-based loss functions
- Optimal experimental design
  - A-optimal design
  - D-optimal design
  - E-optimal design
  - Non-linear optimal experimental design
- Summary

Lyle H Ungar, University of Pennsylvania
Optimal design in non-linear models

- Given a non-linear model \( y = g(x, \theta) \)
- Model is described by a Taylor expansion around a
  - \( a_j(x, \hat{\theta}) = \frac{\partial g(x, \theta)}{\partial \theta_j} \), evaluated at \( \hat{\theta} \)

\[
Y_i = g(x, \hat{\theta}) + (\theta_1 - \hat{\theta}_1)a_1(x, \hat{\theta}) + \ldots + (\theta_k - \hat{\theta}_k)a_k(x, \hat{\theta})
\]

- Maximization of Fisher information matrix is now the same as the linear model
- Yields a locally optimal design, optimal for the particular value of \( \theta \)
- Yields no information on the (lack of) fit of the model

[Atkinson, 1996]
Optimal design in non-linear models

- **Problem**: parameter value $\theta$, used to choose experiments $F$, is unknown
- Three general techniques to address this problem, useful for many possible notions of “gain”
  1) **Sequential experimental design**: iterate between choosing experiment $x$ and updating parameter estimates $\theta$
  2) **Bayesian experimental design**: put a prior distribution on parameter $\theta$, choose a best data $x$
  3) **Maximin experimental design**: assume worst case scenario for parameter $\theta$, choose a best data $x$
Response Surface Methods

- Estimate effects of local changes to the interventions (queries)
  - In particular, estimate how to maximize the response

- Applications:
  - Find optimal conditions for growing cell cultures
  - Develop robust process for chemical manufacturing

- Procedure for maximizing response
  - Given a set of datapoints, interpolate a local surface
    (This local surface is called the “response surface”)
    - Typically use a quadratic polynomial to obtain a Hessian
  - Hill-climb or take Newton step on the response surface to find next \( x \)
  - Use next \( x \) to interpolate subsequent response surface
Response Surface Modeling

Goal: Approximate the function $f(c) = \text{score}(\text{minimize}(c))$

1. Fit a smoothed response surface to the data points
2. Minimize response surface to find new candidate
3. Use method to find nearby local minimum of score function
4. Add candidate to data points
5. Re-fit surface, repeat

[Blum, unpublished]
Summary

• Active learning (sequential)
  – Query by committee
  – Uncertainty sampling
  – Information-based loss functions

• Optimal experimental design
  – A-optimal design
  – D-optimal design
  – E-optimal design

• Non-linear optimal experimental design
  • Sequential experimental design
  • Bayesian experimental design

• Response surface methods

Lyle H Ungar, University of Pennsylvania