

# Active Learning, Experimental Design

CS294 Practical Machine Learning

April 29, 2008

Alex Shyr

*Original Slides by Barbara Engelhardt*

## Motivation

- Unlabeled data abundant, but labels expensive
- Can query the world for labels
- Want to choose which data to label
  - to maximize the “value” of that data to my problem
  - to minimize the “cost” of labeling
- This lecture covers:
  - how to measure the value of data
  - algorithms to choose the data

## Toy Example: threshold function



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: Sequencing genomes

- What genome should be sequenced next?
- Criteria for selection?
- Optimal species to detect phenomena of interest



[McAuliffe et al., 2004]

## Example: collaborative filtering

- Users usually rate only a few movies; ratings “expensive”
- Which movies do you show users to best extrapolate movie preferences?
  - Also known as *questionnaire design*
- Baseline questionnaires:
  - 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]

## Topics for today

- Introduction: Information theory
- Active learning
  - Query by committee
  - Uncertainty sampling
  - Information-based loss functions
- Optimal experimental design
  - A-optimal design
  - D-optimal design
  - E-optimal design
- Sequential experimental design
- Bayesian experimental design
- Maximin experimental design
- Summary

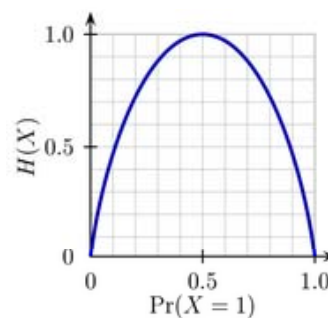
## Topics for today

- Introduction: Information theory
- Active learning
  - Query by committee
  - Uncertainty sampling
  - Information-based loss functions
- Optimal experimental design
  - A-optimal design
  - D-optimal design
  - E-optimal design
- Sequential experimental design
- Bayesian experimental design
- Maximin experimental design
- Summary

## Entropy Function

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

$$H(x) = - \sum_i p(x_i) \log_2 p(x_i)$$
- Comments on entropy function:
  - Entropy of an event is zero when the outcome is known
  - Entropy is maximal when all outcomes are equally likely
- 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  $Q$

$$\begin{aligned}
 D_{\text{KL}}(P \parallel Q) &= \sum_i P(x_i) \log (P(x_i)/Q(x_i)) \\
 &= \sum_i P(x_i) \log P(x_i) - \sum_i P(x_i) \log Q(x_i) \\
 &= H(P, Q) - H(P) \\
 &= \text{Cross-entropy} - \text{entropy}
 \end{aligned}$$

- Measures how different the two distributions are

## KL divergence properties

- Non-negative:  $D(P \parallel Q) \geq 0$
- Divergence 0 if and only if  $P$  and  $Q$  are equal:
  - $D(P \parallel Q) = 0$  iff  $P = Q$
- Non-symmetric:  $D(P \parallel Q) \neq D(Q \parallel P)$
- Does not satisfy triangle inequality
  - $D(P \parallel Q) \not\leq D(P \parallel R) + D(R \parallel Q)$

## KL divergence as gain

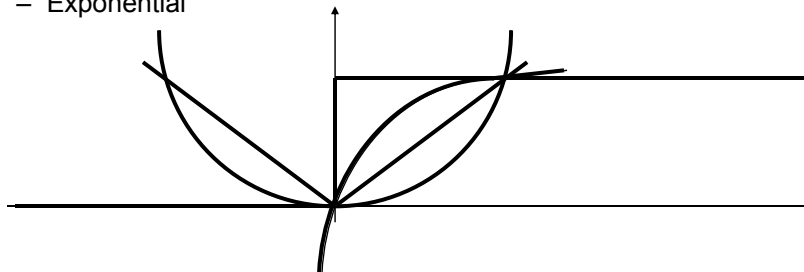
- Modeling the KL divergence of the posteriors measures the amount of information gain expected from query (where  $x'$  is the queried data):

$$D(p(\theta | x, x') || p(\theta | x))$$

- Goal: choose a query that *maximizes* the KL divergence between posterior and prior
- Basic idea: largest KL divergence between updated posterior probability and the current posterior probability represents largest gain

## Loss Functions

- A function  $L$  that maps an event to a real number, representing cost or regret associated with event
- E.g., in regression problems,  $L(y, \theta^T f(x))$  maps to reals
- Examples:
  - Quadratic (least squares) loss
  - Linear (absolute value) loss
  - 0-1 (binary) loss
  - Exponential



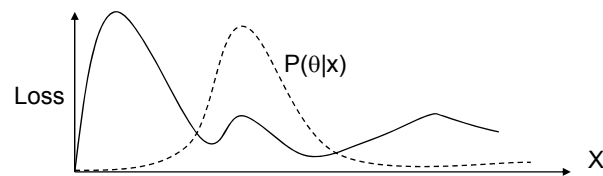
## Risk Function

- *Risk* is also known as *expected loss*
- The (frequentist) risk function is explicitly expected loss

$$R(\Theta, X) = \sum_x L(\theta, x) p(x|\theta)$$

- Bayes risk is defined as posterior expected loss:

$$R(\Theta, X) = \sum_{\theta} L(\theta, x) p(\theta|x)$$

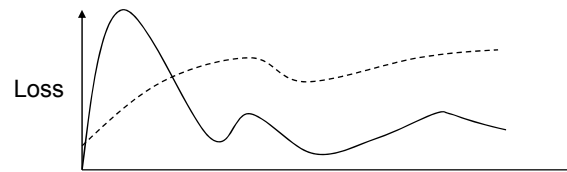


- Goal: choose  $x$  that minimizes expected loss

## Minimax loss

- Wald's (1950) alternative: to minimize the maximum (expected) loss
- Assume response  $x$  is the worst case scenario (gives the greatest expected loss)

$$\text{Minimax}(X, \Theta) = \min_{\theta} \max_x \text{Loss}(X, \theta')$$



- Our problem can be thought of as maximizing the minimum gain (maximin)

## Topics for today

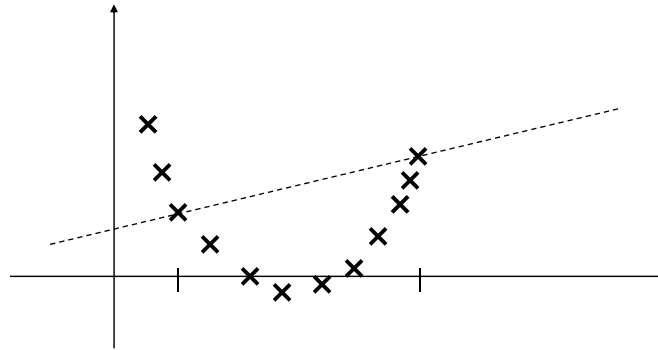
- Introduction: information theory
- Active learning
  - Query by committee
  - Uncertainty sampling
  - Information-based loss functions
- Optimal experimental design
  - A-optimal design
  - D-optimal design
  - E-optimal design
- Sequential experimental design
- Bayesian experimental design
- Maximin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

## What is Active Learning?

- Unlabeled data are readily available; labels are expensive
- Want to use adaptive decisions to choose which labels to acquire for a given dataset
- Goal is accurate classifier with minimum cost (ie. minimum number of examples)

## 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?

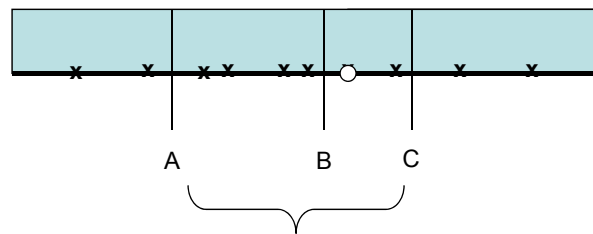


## Active Learning

- *Active learner* is able to query world and receive a response before outputting a classifier
- Learner selects queries (but cannot impact response)
- Two general methods:
  - Select “most uncertain” data given model and parameters
  - Select “most informative” data to optimize expected gain
- Given model  $M$  with parameters  $\theta$  and loss function  $L$
- Query  $q$  with response  $x$  updates the model posterior  $\theta'$

## Query by Committee

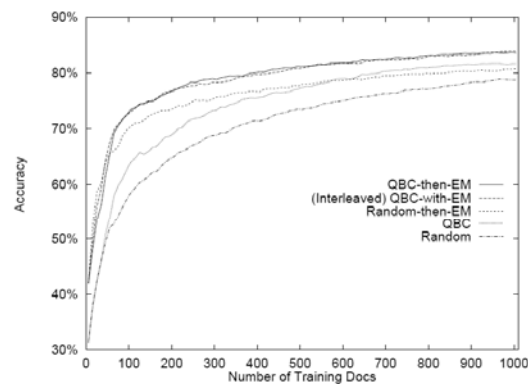
- Queries an example based on the degree of disagreement between committee of classifiers
  - Prior distribution over classifiers/hypotheses
  - Samples a set of classifiers from distribution



[Seung *et al.* 1992, Freund *et al.* 1997]

## Query by Committee Application

- Used naïve Bayes model for text classification in a Bayesian learning setting (20 Newsgroups dataset)

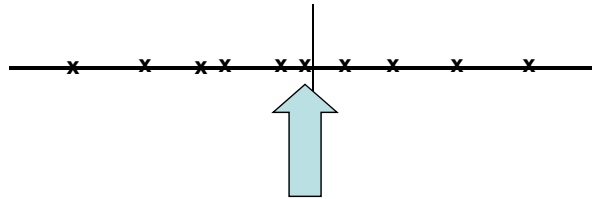


[McCallum & Nigam, 1998]

## Uncertainty Sampling

- Query the event that the current classifier is most uncertain about

If uncertainty is measured in Euclidean distance:



- Used trivially in SVMs, graphical models, etc.

[Lewis & Gale, 1994]

## Information-based Loss Function

- Want to model notions of information gain
  - Maximize **KL divergence** between posterior and prior
  - Maximize reduction in **model entropy** between posterior and prior
  - Minimize **cross-entropy** between posterior and prior
  - Etc.
- All of these can be extended to optimal design algorithms
- Must decide how to handle uncertainty about query response, model parameters

[MacKay, 1992]

Break?

## Topics for today

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

## What is Experimental Design?

- Choose among a menu of possible experiments  $x$

Consider: estimating vector  $z \in \mathbb{R}^n$  from measurements

$$y_i = a_i^T z + \varepsilon_i, \quad i = 1, \dots, m$$

Let  $m_j$  = number of experiments with  $a_i = x_j$ ,  $j = 1, \dots, p$   
and  $\sum m_j = m$

- Each experiment has error rate  $\varepsilon_i$
- For rest of lecture, assume goal is to choose experiments that minimize error covariance matrix
- Under squared error, the error covariance matrix is

$$\left( \sum_j m_j x_j x_j^T \right)^{-1} \quad \text{or} \quad (X^T M X)^{-1}$$

## Optimal Experimental Design Assumptions

- In the setting of Optimal Experimental Design, can assume a few more things
  - Unbiasedness of experiments:  $E\varepsilon_i = 0$
  - Uncorrelatedness of experiments:  $E\varepsilon_i \varepsilon_j = 0$
  - Variance homogeneity:  $E\varepsilon_j^2 = \sigma^2 > 0, j = 1 \dots N$

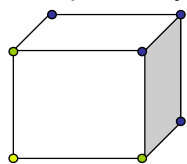
[Atkinson, 1996]

## Experimental Design

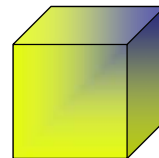
- Select a set of (possibly noisy) queries (or experiments) from  $\mathbf{x}$  that together are maximally informative
- Want to minimize error covariance matrix  $(X^T M X)^{-1}$
- Problem is combinatorial (hard); can be relaxed to convex problem
  - Divide  $M$  by  $m$
  - We will recover the integral solution at the end

## Relaxed Experimental Design

- The *relaxed* problem allows  $w_i \geq 0, \sum_i w_i = 1$
- Error covariance matrix becomes  $(X^T W X)^{-1}$
- $(X^T W X)^{-1}$  = inverted Hessian of the squared error
  - or inverted Fisher information matrix
- minimizing  $(X^T W X)^{-1}$  reduces model error,
  - or equivalently maximize information gain



Boolean problem

 $N = 3$ 

Relaxed problem

## Experimental Design: Types

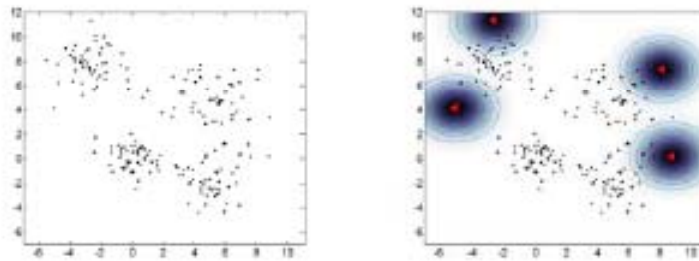
- Want to minimize  $(X^T W X)^{-1}$  ; need a scalar objective
  - *A-optimal* design minimizes the trace of  $(X^T W X)^{-1}$
  - *D-optimal* design minimizes log determinant of  $(X^T W X)^{-1}$
  - *E-optimal* design minimizes maximum eigenvalue of  $(X^T W X)^{-1}$
- All of these design methods can use convex optimization techniques
- Computational complexity polynomial for semi-definite programs (*A-* and *E-optimal* designs)

[Boyd & Vandenberghe, 2004]

## A-Optimal Design

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

Example: mixture of four Gaussians



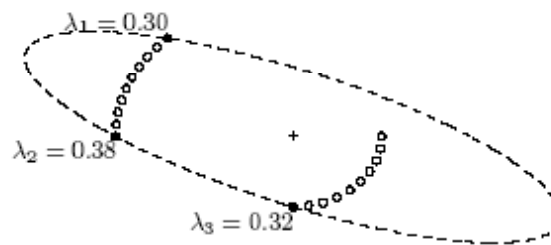
(a) Data set

(b) A-optimal design [Yu et al., 2006]

## A-Optimal Design

- *A-optimal* design minimizes the trace of  $(X^T W X)^{-1}$ 
  - Can be cast as a semi-definite program

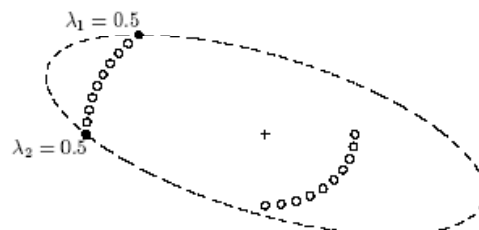
Example: 20 datapoints, minimal ellipsoid that contains all points



[Boyd & Vandenberghe, 2004]

## D-Optimal design

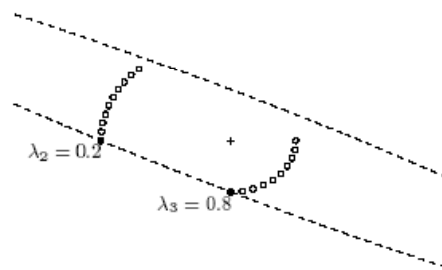
- *D-optimal* design minimizes log determinant of  $(X^T W X)^{-1}$
- Equivalent to choosing the confidence ellipsoid with minimum volume
- Note: non-zero experiment weights are equal



[Boyd & Vandenberghe, 2004]

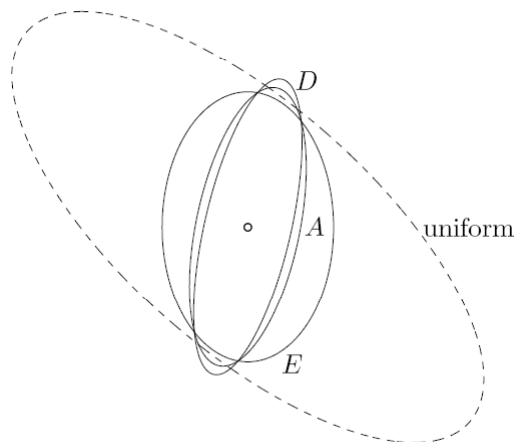
## E-Optimal design

- *E-optimal* design minimizes largest eigenvalue of  $(X^T W X)^{-1}$
- Equivalently, minimizes the norm of  $(X^T W X)^{-1}$ 
  - Can be cast as a semi-definite program
- Minimizes the diameter of the ellipsoid



[Boyd & Vandenberghe, 2004]

## Summary of Optimal Design



[Boyd & Vandenberghe, 2004]

## Optimal Design

- Extract the integral solution for the unrelaxed problem
- Can simply round the weights to closest multiple of  $1/m$ 
  - $m_j = \text{round}(m * w_j), i = 1, \dots, p$

[Boyd & Vandenberghe, 2004]

## Extensions to optimal design

- Cost associated with each experiment
  - Add a cost vector, constrain total cost by a budget  $B$  (one additional constraint)
- Multiple samples from single experiment
  - Each  $x_i$  is now a matrix instead of a vector
  - Optimization (covariance matrix) is identical to before
- Timeline/ordering of experiments
  - Add time dimension to each experiment vector  $x_i$

[Atkinson, 1996]

[Boyd & Vandenberghe, 2004]

## Topics for today

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

## Optimal design in non-linear models

- Given a non-linear model  $y = g(x, \theta)$
- The model is described by a Taylor expansion around a  $\theta_0$ 
  - $a_j(\theta) = \partial g(x, \theta) / \partial \theta_j$ , evaluated at  $\theta_0$
- Maximization of 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  $X$ , is unknown
- Three general techniques to address this problem, useful for many possible notions of “gain”
- **Sequential experimental design**: iterate between choosing experiment  $x$  and updating parameter estimates  $\theta$
- **Bayesian experimental design**: put a prior distribution on parameter  $\theta$ , choose a best data  $x$
- **Maximin experimental design**: assume worst case scenario for parameter  $\theta$ , choose a best data  $x$

## Sequential Experimental Design

- Model parameter values are not known exactly
- Multiple experiments are possible
- Learner assumes that only one experiment is possible; makes best guess as to optimal data point for given  $\theta$
- Each iteration:
  - Select data point to collect via experimental design using  $\theta$
  - Single experiment performed
  - Model parameters  $\theta'$  are updated based on all data  $x'$
- Similar idea to Expectation Maximization

[Pronzato & Thierry, 2000]

## Bayesian Experimental Design

- Effective when knowledge of distribution for  $\theta$  is available
- Example: KL divergence between posterior and prior
  - $\int_x \operatorname{argmax}_w \int_{\theta \in \Theta} D(p(\theta|w,x) || p(\theta)) p(x|w) d\theta dx$
- Example: A-optimal design:
  - $\int_x \operatorname{argmax}_w \int_{\theta \in \Theta} \operatorname{tr}(X^T W X)^{-1} p(\theta|w,x) p(x|w) d\theta dx$
- Often sensitive to distributions

[Chaloner & Verdinelli, 1995]

## Maximin Experimental Design

- Maximize the minimum gain
- Example: D-optimal design:
  - $\operatorname{argmax}_w \min_{\theta \in \Theta} \log \det (X^T W X)^{-1}$
- Example: KL divergence:
  - $\operatorname{argmax}_w \min_{\theta \in \Theta} D(p(\theta|w,x) || p(\theta))$
- Does not require prior/empirical knowledge
- Good when very little is known about distribution of parameter  $\theta$

[Pronzato & Walter, 1988]

## Topics for today

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

## Related ML Problems

- Online Learning, Reinforcement Learning
  - Interaction with the world

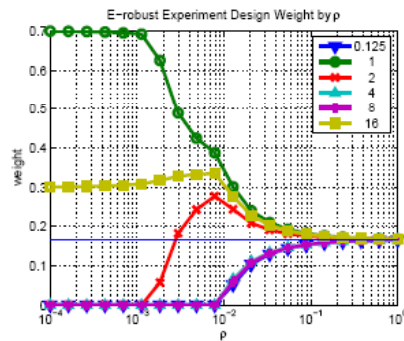
## Summary

- Active learning
  - Query by committee Distribution over parameter;  
Probabilistic; sequential
  - Uncertainty sampling Distribution over parameter;  
Distance function; sequential
  - Information-based loss functions Maximize gain; sequential
  
- Optimal experimental design
  - A-optimal design Minimize trace of information matrix
  - D-optimal design Minimize log det of information matrix
  - E-optimal design Minimize largest eigenvalue of information matrix
- Sequential experimental design Multiple-shot experiments;  
Little known of parameter
- Bayesian experimental design Single-shot experiment;  
Some idea of parameter  
distribution
- Maximin experimental design Single-shot experiment;  
Little known of parameter  
distribution (range known)

## Extra Stuff

## Example: experimental design

- Design experiment to show enzyme reacting with substrate S
- Problem: what concentration(s) of the substrate to test?
- Experiment result is velocity of the reaction, as modeled by a non-linear equation



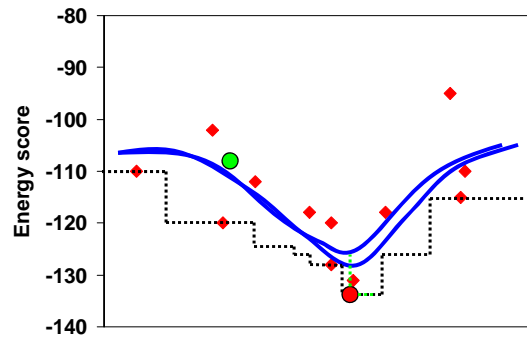
[Flaherty et al. 2005]

## Response Surface Methods

- Estimate effects of and interactions between local changes to the experiments
- Given a set of datapoints, interpolate a local surface (This local surface is called the “response surface”)
- Hill-climb on the response surface to find next  $x$
- Use next  $x$  to interpolate subsequent response surface
- Note: this is original model for which optimal experimental designs were developed [Kiefer, 1959]

## Example: Response Surface

- 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]