Machine Learning Overview

Lyle Ungar

Kinds of machine learning

- **Supervised**
- **Unsupervised**
- **Semi-supervised**
- **Reinforcement**
- **Flavors**
  - Regression vs. classification
  - Parametric vs. nonparametric
  - Active vs. passive
  - Single task vs. multi-task

Supervised learning

- **Non-parametric**
- **Parametric**
  - Minimize error
  - Maximize likelihood (MLE/MAP)
- **‘Semiparametric’**

Bias-Variance tradeoff

Supervised learning

- **Non-parametric**
  - K-NN, Decision Trees, Random Forests, Boosted Trees
- **Parametric**
  - **Regression:** linear, logistic, LMS
  - **Large margin:** SVM, perceptron
- **Semiparametric**
  - neural nets

Loss functions

- **Real y**
  - L_2
  - L_1
- **Categorical y**
  - L_0
  - Hinge
  - Log loss: - $\Sigma_i \, log(p_i)$
    - p_i = the estimated probability of the correct answer
    - minimizes KL(y|p)

Which loss function for classification?

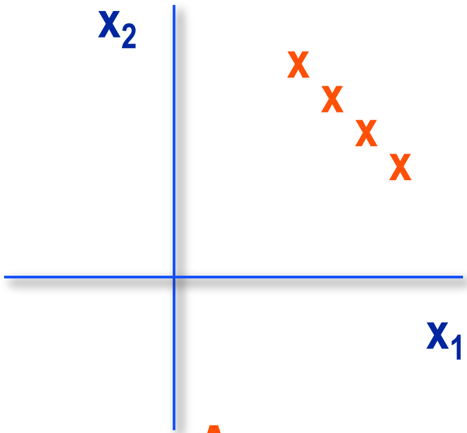
- L_2 vs **log loss**
  - Which is preferred? Why?
- L_0 vs **hinge** vs **log loss**
  - Which is most “hard”?
  - Which is most “soft”?
  - Which fits a probability model?

Unsupervised learning

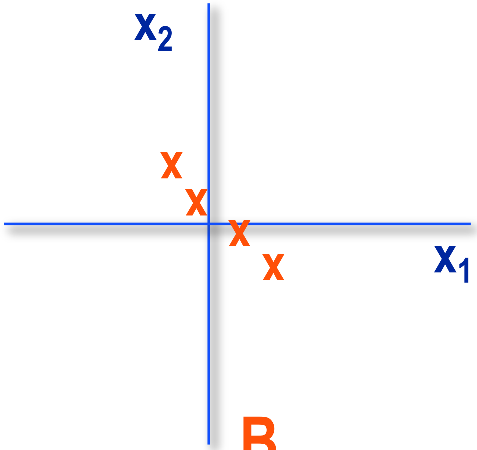
- **Projection** vs. **clustering**
- **Minimize reconstruction error**
  - PCA
  - K-means
  - Auto-encoders
- **Maximize likelihood**
  - Gaussian Mixture Model (GMM)
  - LDA
  - Belief nets, including Naïve Bayes

When to mean center for PCA?

- Product purchases (e.g. amazon)
- Word counts (e.g. twitter)
- Pixels (e.g. brain scans)



A



B

When (not) to rescale

- OLS
- Ridge, elastic net
- K-NN
- RBF
- PCR
- SVM
- Convolutional neural net
- Random forest, boosted trees

Scale invariant?

Method Selection: How big is n vs p ?

- $p \gg n$
- $n \gg p$
- $n \sim p$

Method Selection: How big is *n* vs *p* ?

- ***p* >> *n*:** **use dimensionality reduction**
  - or do extreme feature selection (RIC)
  - Then often just fit a linear model
  - Try semi-supervised learning
- ***n* >> *p*:** **fit a flexible model**
  - random forest, NNet, boosted trees
  - or look for more features
- ***n* ~ *p*:** **consider feature selection – and dim. reduction**
  - Elastic net?

What do you know about your problem?

- Are features highly correlated or almost independent?
- Roughly linear or highly nonlinear?
- Is noise Gaussian?
- Conditional independence or causal structure?
- Constraints?
- Fixed size or variable length feature set?
- What is your real loss function?

What method to use? Why?

Data	#y classes	n	p
◆ MRI	2	100	10,000
◆ Image	1,000	500,000	600
◆ Disease	3	1,000	50
◆ Disease	10	1,000	200
◆ Text in docs	2	40,000	40,000
◆ Student apps	2	5,000	500