Course Outline

1. Introduction to the course and sample crawling task
2. Classification, hands on with Weka
3. K-means, Topic modeling, demo with Mallet
4. PageRank, Gephi demo
5. Information extraction, OpenCalais demo

Each lecture: concepts + specific tasks in CiteSeer + demo/exercise

Course homepage

http://www.cse.unt.edu/~ccaragea/russir14/schedule.html

Classification Problems

- Email filtering: spam / non spam
- Email foldering/tagging: Work, Friends, Family, Hobby
- Research articles by topics: Machine Learning, Data Mining, Algorithms
- Document by type: research article, thesis, slides, CV
- Tumor: malignant / benign
- Medical diagnosis: Not ill, Cold, Flu

Assign each document to a label from a known set of labels

We have a labeled dataset (*supervised* learning)

Classification algorithms

- Tree-based models: automatically generate conjunctive rules
- Generative models:
  - Estimate probability distributions for data and apply Bayes' theorem

    1. Assume a generative distribution for data
    2. Estimate parameters for class priors and data class distributions from the training data
    3. Use posterior probabilities for prediction

Estimate parameters of the distribution

- Very specific to the form of the assumed distribution
- Maximum likelihood estimate

$$p(\vartheta|\mathcal{X}) = \frac{p(\mathcal{X}|\vartheta) \cdot p(\vartheta)}{p(\mathcal{X})}, \qquad \text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}.$$

$$L(\vartheta|\mathcal{X}) \triangleq p(\mathcal{X}|\vartheta) = \bigcap_{x \in \mathcal{X}} \{X = x|\vartheta\} = \prod_{x \in \mathcal{X}} p(x|\vartheta),$$

$$\hat{\vartheta}_{\mathrm{ML}} = \operatorname*{argmax}_{\vartheta} \mathcal{L}(\vartheta|\mathcal{X}) = \operatorname*{argmax}_{\vartheta} \sum_{x \in \mathcal{X}} \log p(x|\vartheta). \qquad \frac{\partial \mathcal{L}(\vartheta|\mathcal{X})}{\partial \vartheta_k} \stackrel{!}{=} 0 \quad \forall \vartheta_k \in \vartheta.$$

Example

Bernoulli density function (p: probability of throwing a head, c=0/1)

$$p(C{=}c|p) = p^c\ (1-p)^{1-c} \triangleq \mathrm{Bern}(c|p)$$

$$\begin{aligned}\mathcal{L} = \log \prod_{i=1}^{N} p(C{=}c_i|p) &= \sum_{i=1}^{N} \log p(C{=}c_i|p) \\ &= n^{(1)} \log p(C{=}1|p) + n^{(0)} \log p(C{=}0|p) \\ &= n^{(1)} \log p + n^{(0)} \log(1-p)\end{aligned}$$

where $n^{(c)}$ is the number of times a Bernoulli experiment yielded event c. Differentiating with respect to (w.r.t.) the parameter p yields:

Example (contd.)

$$\frac{\partial \mathcal{L}}{\partial p} = \frac{n^{(1)}}{p} - \frac{n^{(0)}}{1-p} \stackrel{!}{=} 0 \quad \Leftrightarrow \quad \hat{p}_{\mathrm{ML}} = \frac{n^{(1)}}{n^{(1)} + n^{(0)}} = \frac{n^{(1)}}{N},$$

which is simply the ratio of heads results to the total number of samples.

- Avoid zero probabilities

    Fold in priors and prior distributions with hyperparameters

        MAP estimates, Bayesian estimates

Naive Bayes Multinomial for text

- Assume a generative distribution
  - Each class has a multinomial distribution over terms
- Compute parameters based on training data

- Calculate $P(c_j)$ terms
  - For each c_j in C do
    $docs_j \leftarrow$ all docs with class $=c_j$
    $$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$
- Calculate $P(w_k \mid c_j)$ terms
  - $Text_j \leftarrow$ single doc containing all $docs_j$
  - For each word w_k in $Vocabulary$
    $n_k \leftarrow$ # of occurrences of w_k in $Text_j$
    $$P(w_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

- Use posteriors for assigning class labels
- Bag-of-words and conditional independence assumptions

Discriminative Models

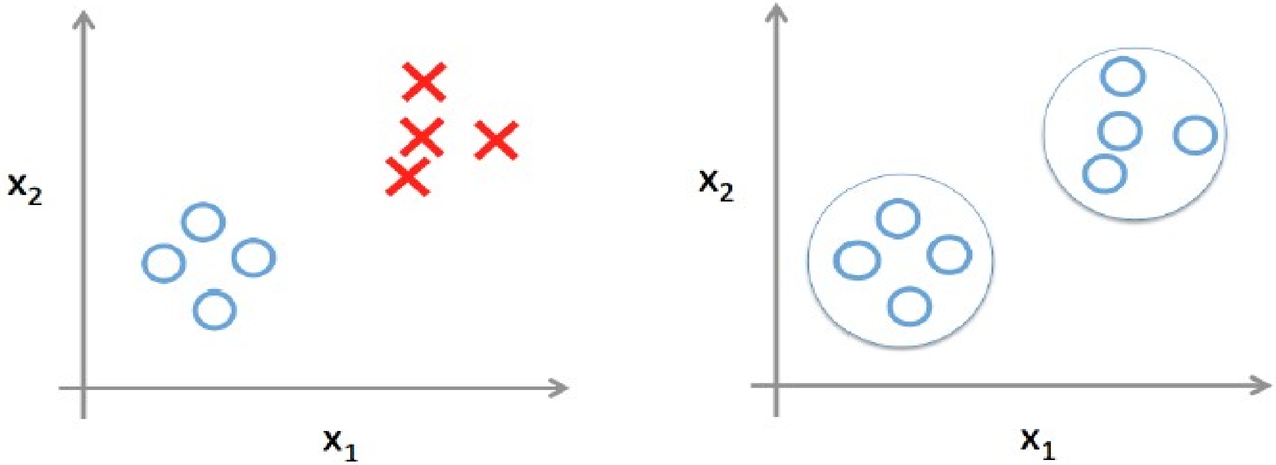
- Don't care for generating the data P(x,c) but instead model conditional directly P(c|x)
  - Maximum Entropy P(c|x) is f(w_c.x)

- Usually talk in terms of weight/parameter vectors for features
- Computing parameters based on training data
  - Involves numerical optimization of a loss function on the training data

Questions from yesterday

- Decision trees over unsupervised data
  - D. Karakos, S. Khudanpur, J. Eisner and C. E. Priebe, Unsupervised Classification via Decision Trees: An Information-Theoretic Perspective, in Proceedings of the 2005 IEEE International Conference on Acoustics, Speech and Signal Processing
- Basic/introductory course on ML
  - Several courses on coursera (Andrew Ng's course)

Unsupervised Learning



Supervised learning　　　Unsupervised learning

- **Unsupervised learning:** Learning to group objects into categories, without any training labels.
  - Examples: clustering search results into topics

Popular approaches

- Clustering
  - K-means
  - Hierarchical clustering
  - Graph-based clustering
  - Density-based clustering, DBSCAN
- Mixture models
  - Topic modeling
  - EM-based models
- Dimension Reduction
  - Principal Component Analysis
  - Matrix factorization