Final Review

Monday, June 06, 2016 4:22 PM

Index:

From Before:

Text Preprocessing

Statistical Properties and Evaluation

Boolean Retrieval

Vector Space Model

Relevance Feedback

Statistical Language Modeling

Document Prior

Page Rank

New:

Page Rank

Recommendation Systems

Text Classification

Text Clustering

Page Rank

Recommendation Systems

Algorithms

- Collaborative Filtering
- Content-Based
- Hybrid

A Collaborative Filtering

- Like users rate similarly

- k-nearest
  - choose the k most similar

- Neighborhood Selection
  Cosine sim

$$w_{a,u} = \frac{\sum_{i=1}^{n} r_{u_1,i} \times r_{u_2,i}}{\sqrt{\sum_{i=1}^{n} r_{u_1,i}^2} \times \sqrt{\sum_{i=1}^{n} r_{u_2,i}^2}}$$

Evaluating Predictions: RMSE

$$\sqrt{\frac{1}{S}\sum_{(u,i)\in test} (p_{u,i} - r_{u,i})^2}$$

- no under/over

p ratings predicted
r true rating
(u,i) ∈ test missing ratings
S total predicted

Rating Prediction

$$p_{a,i} = \frac{\sum_{u=1}^{k} w_{a,u} r_{u,i}}{\sum_{u=1}^{k} w_{a,u}}$$

- Pearson Correlation
  - model the deviation

$$w_{a,u} = \frac{\sum_{i=1}^{t} (r_{a,i} - \bar{r}_a)(r_{u,i} - \bar{r}_u)}{\sqrt{\sum_{i=1}^{t} (r_{a,i} - \bar{r}_a)^2} \times \sqrt{\sum_{i=1}^{t} (r_{u,i} - \bar{r}_u)^2}}$$

$\bar{r}$ average

sub results:
1 - perfectly correlated
0 - not
-1 - inverse

$$p_{a,i} = \bar{r}_a + \frac{\sum_{u=1}^{k} w_{a,u}(r_{u,i} - \bar{r}_u)}{\sum_{u=1}^{k} |w_{a,u}|}$$

Optimizations: Improving Predictions
- Penalize universally liked movies

$$IUF(j) = \log \frac{m}{m_j}$$

m — total # users
m_j — # users rated item j

- multiply original ratings by IUF during weight calc

- Case Amplification

$$w_{a,u} = w_{a,u} \cdot |w_{a,u}|^{\rho-1}, \quad \rho = 2.5, \geq 1$$

- favors high weights, punishes low weights

Euclidean dist:

$$d = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}, \quad w = \frac{1}{d+1}$$

Jaccard Sim

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

→ binary ratings

User Based - Summary
1) find k nearest
2) use ratings from k-nearest to predict for the active user

Item-Based Collaborative Filtering
- Similar Items rated similarly - reverse users & items in previous

Cos sim

$$w = \sum r_{u_j,i_1} \times r_{u_j,i_2}$$

Pearson: $\frac{\sum (R_{u,i} - \bar{R}_i)(R_{u,j} - \bar{R}_j)}{}$

$$\sqrt{\sum r_{u,i_1}^2} \times \sqrt{\sum r_{u,i_2}^2}$$

$$\sqrt{\sum (R_{u,i} - \bar{R}_i)^2} \sqrt{\sum (R_{u,j} - \bar{R}_j)^2}$$

Adj cos sim ↓ Replace $\bar{R}_i$ w/ $\bar{R}_u$

User vs Item

- Item more efficient
- Item more stable - users have multiple interests
- Item based - little diversity or surprise - obvious and boring

Cold-Start Problem

- new items/users have no historical ratings

Content-Based

content compare to user profile

IR applies - treat user as a query, item as document

K-Nearest - Probability improvements

- center your data
- more points less noisy, orders of magnitude for ratings
- Smoothing!

Linear - $r_u = \alpha \cdot r_u + (1-\alpha) \cdot g$ → g global mean $\quad r_u$ is r_u average rating

r_u $\frac{\alpha \text{ fixed}, \text{user ratings to apply}}{\beta + n_u} \cdot r_u + \frac{}{\beta + n_u} \cdot g$ → Dirichlet $\quad$) n_u # ratings of user u

- get a shrunken mean

~~Text Classification~~

- Assign Categories to text documents
- Vector Space - word = component

Like Rocchio

$$\mu_c = \frac{1}{|D_c|} \sum_{d \in D_c} \bar{v}(d)$$

D_c set of all documents in class c, $v(d)$ vector rep of d

K-nearest

- keep all training docs
- k docs that are most similar to the new doc
- assign category that is most common amongst neighbor docs

$$\hat{y}(x) = \frac{1}{\sum (\text{sim}(x, x_i))} \sum \text{sim}(x, x_i) y_i$$

$$\begin{cases} 1 & \text{if } \hat{y}(x) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

similarity functions

- euclidean, KL divergence, Dot product, cosine similarity

Naive Bayes Text Classification

- query likelihood, doc prior

C - doc of all docs in

the class

$P(C,d) = P(C|d)P(d) = P(d|C)P(C)$ the prior

$P(C|D) = \frac{P(d|C)P(C)}{P(d)}$

d - query
- compute likelihood

$C = \arg\max P(C_j) \prod_{i}^{n} P(t_i|C_j)$

$P(t_i|C_j) = \frac{tf(t_i, C_j)}{|C_j|}$, $P(C_j) = \frac{\#\text{docs in } C_j}{\#\text{docs}}$

smoothing $\Rightarrow$

add-one, etc

$P(t_i|C_j) = \frac{tf(t_i, C_j) + 1}{|C_j| + |V|}$

	in the class	not in the class
predicted to be in the class	true positives (TP)	false positives (FP)
predicted to not be in the class	false negatives (FN)	true negatives (TN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F = 2 Precision * Recall/(Precision + Recall)*

Accuracy = (TP + TN)/(TP+FP+FN+TN)

Text Clustering

- unsupervised learning
- inferred from data

vs classification
supervised
user defined

☆ K-means

$\vec{\mu}(c) = \frac{1}{|c|} \sum_{\vec{x} \in c} \vec{x}$

- iterative - based on centroids
- reassignment based on distance to current cluster centroids
- Recompute centroids based on new membership

- guaranteed to converge
- not optimal - need good seeds
- only on linear data

$k \simeq \sqrt{n/2}$, n # data points

Final Sample

1) Summarize algorithms

a) User-based. Like users rate similarly

b) Item-based. Similar items are rated similarly

2) Major disadvantage to RMSE?

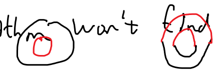
Subjective Metric ≠ RMSE
- doesn't correlate w/ user-satisfaction

Can't differentiate between above & below ground truth
- no under/over

3) K-means doesn't work where?

Non-gaussian data — ex: concentric circles.

The centroids are the same, but the algorithm won't find that



vs

4) C = America, $\bar{C}$ = everything else.

5 training, 1 test : 6 total docs Add-one smoothing

$\frac{P(d|C)\,P(C)}{P(d)} \propto P(d|C)\,P(C)$

$P(\text{America}) = 5/12$ ← 11 total words + 1 $= 5/12$

$P(\text{Cali}) = 1/12$

$P(\text{Wash}) = 1/12$

$P(\text{oregon}) = 1/12$

$P(\text{Tokyo}) = 2/12$

$P(\text{Japan}) = 1/12$

$P(C) = (5/12)^3 \cdot (1/12) \cdot (1/12) \cdot 3/4$

$P(C) = 3/4$ ← doc prior — 3/4

$P(\bar{C}) = 1/4$ ← 1 doc not in C

~~$P(t_i|c_j) = \frac{T(t_i, c_j) + 1}{|c_j| + |V|}$~~

$P(d|C) = \prod P(t|C)$, $t \in d$ $[P(C)]$

$= P(\text{America}|C)^3 \cdot P(\text{Tokyo}|C) \cdot P(\text{Japan}|C)$

$\Rightarrow |c_j| = 7$, $|V| = 6$

$= \left[\frac{5}{13}\right]^3 \left[\frac{1}{13}\right] \left[1/13\right] \left[3/4\right] = 2.52e^{-4}$

$P(d|\bar{C}) = \left[\frac{2}{7}\right]^3 \left[\frac{3}{7}\right] \left[\frac{2}{7}\right] \left[1/4\right] = 7.13\,e^{-4}$

$\bar{c_j} = 4$ $|V| = 3$

$P(d|\bar{C}) > P(d|C) \Rightarrow$ in $\bar{C}$