Internet Analytics (COM-308)

Homework Set 4

Exercise 1

You are training a machine learning model on some training data, and then evaluate the error on some separate validation data that you kept aside. You notice that the validation error is considerably larger than the training error. Is this normal, or is there a problem? If there's a problem, how would you fix it by adjusting the regularizer weight λ? How will the two errors evolve if you change λ?

Exercise 2

The rating matrix of five users for four items with few missing entries is given by

$$R = \begin{bmatrix} 2 & 1 & 3 & - \\ 3 & 2 & 5 & 5 \\ 5 & - & 4 & 2 \\ 4 & 3 & - & 4 \\ - & 1 & 5 & 3 \end{bmatrix}.$$

(a) First we try a simple predictor that takes into account only user bias. Compute the optimal baseline predictor $\overline{r} + b_u$ (without regularization) and the RMSE for this predictor. Feel free to rely on a tool such as matlab, octave, or mathpy to solve the resulting system of equations.

(b) Now we refine this predictor to take into account both a user and an item bias, as seen in class. Compute the optimal baseline predictor $\overline{r} + b_u + b_i$ (without regularization) and the RMSE for this predictor.

(c) Predict the missing values in the rating matrix R for the predictors in parts (a) and (b).

Exercise 3

(a) The most basic version of stochastic gradient descent (SGD) works as follows. There are n data points $(x_1, \ldots, x_n)$. We want to minimize a function $f(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} f_i(\theta; x_i)$ with respect to θ (the model parameters to be optimized).

Instead of performing gradient descent using the full gradient $\nabla_\theta f$, we select, for each step in the iteration, a data index $I \sim \text{unif}(1, n)$ randomly, and we update the current estimate of θ with the gradient $\nabla_\theta f_I$.

What is the expected gradient $\mathrm{E}[\nabla_\theta f_I]$?

(b) As we have seen in class, training the recommender system model on data amounts to finding

$$(P^*, Q^*) = \arg\min_{P,Q} \sum_{(u,i) \in R} (r_{ui} - p_u^T q_i)^2 + \lambda(\|P\|^2 + \|Q\|^2).$$

Compute the gradient $\nabla_{P,Q} f(P, Q; x_1, \ldots, x_n)$ used in gradient descent for this model.

Exercise 4

Assume you visit an on-line news website and you want to read about sports and do not like reading about politics. You use naïve Bayesian text classification to select interesting headers. Suppose you have the following training corpus of news headlines of sports (G) and politics (B):

Sports (the headers you like)	Politics (the headers you would like to filter out):
"victory Switzerland against US "	"US elections victory "
"worldcup finals results "	"results campaign against party"
"powerful victory against Switzerland"	"US cancel elections results"

Using Laplace smoothing (with $k = 1$) compute the smoothed priors $P(B), P(G)$, and the smoothed word model $P(W|G), P(W|B)$, where W is a word.

W	$P(W\|G)$	$P(W\|B)$
cancel		
powerful		
against		
US		
worldcup		
finals		
results		
Switzerland		
elections		
campaign		
party		
victory		

Compute the posterior probabilities $P(G|M), P(B|M)$ for the following messages M and classify them as B or G:

1. M_1="worldcup victory "
2. M_2="victory against US"
3. M_3="US campaign results. Victory party: .."

Suppose a political-news writer tries to manipulate his actual header M_3 by adding words with a purpose of getting through your filter.

- Which dummy word added to M_3 (resulting in M_3') increases the posterior $P(G|M_3')$ the most?
- How many times would he need to add this word to M_3 to classify the header as sports instead of politics?