INF1771 - INTELIGÊNCIA ARTIFICIAL

TRABALHO 3 – APRENDIZADO DE MÁQUINA

Descrição:

Redes sociais tem se tornado extremamente populares nos últimos anos. Facebook, Twitter, Google+ e **Tencent Weibo** recebem milhares de novos usuários a cada dia. Tencent Weibo é uma rede social pouco conhecida no Brasil, mas é uma das redes sociais mais utilizadas na China. Desde o seu lançamento em abril de 2010, o Tencent Weibo se tornou uma plataforma importante para a construção de amizades e compartilhamento de interesses online. Atualmente, existem mais de 200 milhões de usuários registrados no Tencent Weibo, gerando mais de 40 milhões de mensagens por dia.

Redes sociais utilizam ativamente varias técnicas de aprendizado de máquina, por exemplo, os **sistemas de recomendação**. Um sistema de recomendação tem como objetivo selecionar itens personalizados com base nos interesses dos usuários. Tais itens podem assumir formas bem variadas como, por exemplo, livros, filmes, notícias, música, vídeos, anúncios, links patrocinados, páginas de internet, etc. Empresas como Amazon, Netflix e Google são reconhecidas pelo uso intensivo de sistemas de recomendação com os quais obtém grande vantagem competitiva.

A captura de interesses dos usuários é uma característica fundamental e crucial de sites de redes sociais. Permitindo que o serviço indique aos seus usuários itens potencialmente interessantes (por exemplo, amigos, notícias, jogos, propagandas, produtos), evitando também a indicação de itens que não despertam nenhum interesse dos usuários.

O **objetivo do Trabalho 3** é desenvolver um sistema de recomendação para redes sociais que preveja se um usuário ira ou não seguir um determinado item recomendado pelo sistema. Estes itens podem ser pessoas, organizações ou grupos. Para essa tarefa serão utilizados os dados fornecidos pela rede social **Tencent Weibo**.

Este trabalho é baseado no Track 1 da **KDD Cup 2012** (http://www.kddcup2012.org/). A KDD Cup é uma competição anual de aprendizado de máquina e data mining. A competição está acontecendo nesse momento. Se alguém se interessar em participar tem até 01/06 para enviar a sua submissão, o prêmio para o primeiro colocado é $5.000.

Dataset:

O conjunto de dados (dataset) que será utilizado neste trabalho é constituído de amostras das preferencias dos usuários do Tencent Weibo para diversos itens. Ou seja, os itens recomendados para os usuários e o histórico indicando se o usuário aceitou ou não a recomendação. Além disso, o dataset também fornece informações detalhas sobre o perfil dos usuários e também uma estrutura de categorização dos itens.

Todas as informações do dataset são anônimas, ou seja, os usuários são representados por um identificador numérico, assim como os itens, categorias, palavras-chaves e outras informações.

Os dataset é formado pelos seguintes arquivos:

1) **rec_log_train.txt**

   Descrição: Relaciona usuários com itens sugeridos indicando se o usuário aceitou ou não a sugestão.

   Formato: (*UserId*)\t(*ItemId*)\t(*Result*)\t(*Unix-timestamp*)

   - *UserId*: representa um identificador único do usuário.
   - *ItemId*: representa um identificador único do item.
   - *Result*: valor 1 ou -1 indicando se o usuário aceitou a recomendação (1) ou não aceitou a recomendação (-1).
   - *Unix-timestamp*: timestamp (unix) de quando o evento aconteceu.

   Exemplo: 1529353 1774509 -1 1318348786

2) **user_profile.txt**

   Descrição: Informações referentes aos usuários.

   Formato: (*UserId*)\t(*Year-of-birth*)\t(*Gender*)\t(*Number-of-tweet*)\t(*Tag-Ids*)

   - *UserId*: representa o identificador único do usuário.
   - *Year-of-birth*: indica o ano de nascimento do usuário.
   - *Gender*: sexo do usuário – desconhecido (0), masculino (1) ou feminino (2).
   - *Number-of-tweet*: número inteiro representando a quantidade de mensagens que o usuário já postou na rede social.
   - *Tag-Ids*: Tags selecionadas pelo usuário para descrever o seus interesses. Se o usuário gostar de escalar montanhas e nadar ele provavelmente selecionará “escalar montanhas” e “nadar” como tags para o seu perfil. As tags em

linguagem natural não estão presentes, cada palavra foi substituída por um identificador numérico único. As tags são apresentadas no seguinte formato:

"*tag-id$_1$;tag-id$_2$;...;tag-id$_N$*"

Se o usuário não definiu nenhuma tag para o seu perfil, o valor desse atributo será 0.

Exemplo: 100676 1987 1 8 92;35;41

3) item.txt

Descrição: Informações referentes aos itens.

Formato: (*ItemId*)\t(*Item-Category*)\t(*Item-Keyword*)

- *ItemId*: representa o identificador único do item.
- *Item-Category*: string "a.b.c.d" onde as categorias na hierarquia são delimitadas por "." e ordenados de forma top-down (categoria "a" é pai de "b", "b" é pai de "c" e assim por diante).
- *Item-Keyword*: contém as palavras-chaves extraídas do perfil do item. As palavras-chaves em linguagem natural não estão presentes, cada palavra foi substituída por um identificador numérico único. As palavras-chaves são apresentadas no seguinte formato:

"*id$_1$;id$_2$;...;tid$_N$*"

Exemplo: 862829 1.6.2.1 3824;5273;5300;3824;6183

4) user_action.txt

Descrição: Informações referentes a ações realizadas pelos usuários.

Formato:(*UserId*)\t(*Action-Destination-UserId*)\t(*Number-of-at-action*)\t (*Number-of-retweet*)\t(*Number-of-comment*)

- *UserId*: representa o identificador único do usuário.
- *Action-Destination-UserId*: identificador do usuário ao qual *UserId* estava se referenciando.
- *Number-of-at-action*: número inteiro representando a quantidade de vezes que *UserId* postou uma mensagem referenciando *Action-Destination-UserId*.

- *Number-of-retweet*: número inteiro representando a quantidade de vezes que *UserId* compartilhou uma mensagem criada por *Action-Destination-UserId*.
- *Number-of-comment*: número inteiro representando a quantidade de vezes que *UserId* comentou em uma mensagem criada por *Action-Destination-UserId*.

Exemplo: 1000023 1038725 0 6 0

5) user_sns.txt

Descrição: Informações referentes aos usuários seguidos pelos outros usuários.

Formato: (*Follower-userid*)\t(*Followee-userid*)

- *Follower-userid*: representa o identificador único do usuário que seguiu o usuário *Followee-userid*.
- *Followee-userid*: representa o identificador único do usuário que foi seguido por *Follower-userid*.

Exemplo: 1000004 1774589

6) user_key_word.txt

Descrição: Contém palavras-chaves extraídas das postagens e comentários feitos pelos usuários.

Formato: (*UserId*)\t(*Keywords*)

- *UserId*: representa o identificador único do usuário.
- *Keywords*: contém as palavras-chaves extraídas das postagens e comentários feitos pelo usuário. As palavras-chaves em linguagem natural não estão presentes, cada palavra foi substituída por um identificador numérico único. As palavras-chaves são apresentadas no seguinte formato:

"$keyword_1$:$weight_1$; $keyword_2$:$weight_2$;... $keyword_N$:$weight_N$"

Quanto maior o peso (*weight*), mais interessado o usuário é na palavra-chave.

Exemplo: 1000032 194:0.8508;659:0.4441;1:0.2809

Informações Adicionais:

Para desenvolver o sistema de recomendação você deve utilizar algum método de **aprendizado de máquina supervisionado**, visto que temos um conjunto rotulado de exemplos para treinamento. Não é necessário desenvolver um sistema completo, apenas o núcleo inteligente deste sistema, que neste caso seria o classificador treinado para classificar se um usuário ira ou não aceitar uma sugestão feita pelo sistema.

As tarefas deste trabalho são:

1) Selecionar e estruturar os atributos que serão utilizados para descrever os exemplos de treinamento.
2) Selecionar o algoritmo que será utilizado pelo classificador. A sua escolha obrigatoriamente deve ser justificada com base em testes comparativos realizados com os outros algoritmos disponíveis.
3) Realizar testes excluindo/adicionando atributos aos exemplos, alterando parâmetros dos algoritmos ou realizando outras otimizações. Ou seja, a etapa (1) e (2) devem ser repetidas varias vezes buscando melhorar a precisão da classificação.

Outras Informações Importantes:

- Não é necessário implementar todos os classificadores para realizar os experimentos. É permitida a utilização de bibliotecas externas, como por exemplo, a LibSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/), ou o Weka (http://www.cs.waikato.ac.nz/ml/weka/).

- O conjunto de dados utilizado nesse trabalho está sendo utilizado no Track 1 da KDD Cup 2012. Dessa forma, a página do concurso é uma **importante fonte de informações** para ajudar no entendimento dos dados e elaboração de uma metodologia de modelagem de atributos: http://www.kddcup2012.org

- O conjunto dados que será utilizado é bem grande (**mais de 3 GB de texto**). Você vai precisar desenvolver vários pequenos programas para manipular esses arquivos e gerar as suas bases de treinamento e testes.

- O download dos arquivos com os dados pode ser feito diretamente do site da KDD Cup (http://www.kddcup2012.org/c/kddcup2012-track1/data) ou copiados durante as próximas aulas (tragam um pen drive com espaço livre).

- No KDD Cup existe um conjunto de exemplos de teste. Entretanto, como o concurso ainda está acontecendo, a correta classificação destes exemplos ainda não foi divulgada. Provavelmente essa informação será divulgada após o final

do concurso, se isso acontecer antes do prazo de entrega deste trabalho podemos usar esse conjunto de dados para testes. Antes disso, você deve dividir os dados de treinamento, parte para treinamento e parte para testes.

- A **divisão os dados em treinamento e teste** deve ser feita de forma aleatória. Normalmente 70% dos dados devem ser utilizados para treinamento e 30% para testes. Você deve garantir também que os exemplos de cada classe sejam distribuídos nestas mesmas proporções nas duas bases de dados.

- Tome cuidado para não ser enganado por uma **falsa taxa de acertos** durante o processo de avaliação dos resultados. Existe uma quantidade muito grande de exemplos da classe referente ao usuário não aceitar a recomendação. Sendo assim, se você sempre "chutar" que o usuário "não aceita" você vai ter uma alta porcentagem de acerto geral, mas vai ter errado todas as respostas da outra classe. Para evitar isso você deve calcular o precision e recall (http://en.wikipedia.org/wiki/Precision_and_recall).

Forma de Avaliação:

A avaliação deste trabalho será baseada em um relatório que deverá ser elaborado, entregue e apresentado. Neste trabalho, o mais importante será o **relatório final** e a **apresentação dos resultados**.

O relatório deverá conter:

- Descrição da modelagem dos exemplos de treinamento:
  - Atributos selecionados para descrever os exemplos;
  - Justificativa para a escolha dos atributos;
  - Estrutura dos exemplos;
- Descrição dos experimentos realizados:
  - Variações na modelagem dos exemplos;
  - Variações no conjunto treinamento e testes;
  - Variação nos parâmetros dos algoritmos;
- Comparação dos algoritmos analisados:
  - Taxa de reconhecimento;
  - Tempo gasto no processo de treinamento;
  - Tempo gasto no processo de classificação de um exemplo desconhecido;
- Resultados Finais:
  - Escolha do melhor classificador e conjunto de atributos;
  - Conclusões finais;

Bônus:

- Os trabalhos que conseguirem uma taxa de reconhecimento superior a X% receberam 2.0 pontos extras na nota. Podendo tirar até 12.0 no trabalho. O valor de X ainda será definido.

- Os trabalhos que implementarem um pequeno programa (em qualquer linguagem) que receba como entrada o ID de um usuário, o ID de um item e retorne a indicação se o usuário ira ou não aceitar a recomendação – usando para isso o classificador que obteve melhores nos testes – receberam 0.5 pontos extras na nota.

Data de Entrega:

25/06

Forma de Entrega:

Os trabalhos devem ser **apresentados** na aula do dia 25/06 (segunda) ou 27/06 (quarta).

Todos devem enviar o relatório até o dia 27/06 para o email edirlei.slima@gmail.com.

Não serão aceitos trabalhos enviados depois desta data.