MAC0460 - Lista 3

Parte 1

1. Estudamos o problema de regressão linear e vimos formas de solucionar esse problema. Explique como poderia ser resolvido um problema de regressão polinomial.

   R.: Poderia ser resolvido de maneira similar ao problema da regresssão linear. Ao invés de analisarmos uma relação linear entre x e y, vamos analisar uma relação polinomial. Estaríamos, então, interessados em minizar uma função de erro como:

   $$E(\vec{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(h(x^{(i)}) - y)^2$$

   E a função hipótese é dada por:

   $$h(\vec{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2^2 + \theta_3 x_3^3 + \ldots \theta_n + x_n^n$$

   Ao minimizar a função de erro por meio de algum otimizador (Gradient Descent, por exemplo), obteremos um vetor de parâmetros theta.

---

2. Considerando o contexto de problemas de classificação binária, discuta similaridades e diferenças entre o algoritmo de regressão logística e o algoritmo SVM.

   R.: Podemos ressaltar como similaridade que ambos algoritmos são capazes de criar uma superfície de separação linear nos dados, ou seja, uma reta. Como diferença temos a capacidade do SVM gerar superficies de separação de maior dimensionalidade, por exemplo, em R^2 o SVM consegue traçar curvas complexas para separar os dados. Além disso, o funcionamento dos dois algoritmos diferentes: enquanto que a regressão logística ajusta uma reta separando os dados minimizando o erro quadratico medico, o algoritmo SVM funciona traçando diversas separações nos dados e escolhe aquele modelo com a maior margem. Com isso, geralmente o SVM possui um desempenho mais satisfatório do que regressão logística

---

3. Na sua opinião, confrontando-se redes neurais e SVM, quais são as qualidades interessantes em cada um deles? Explique.

   R.: Uma vantagem que **ambas** possuem é a capacidade de construir hiperplanos de separação complexos: SVMs por meio dos Kernels e Redes Neurais por meio das diversas camadas com não-linearidades. Uma vantagem do **SVM** é que o algoritmo demanda menos dados para fazer uma predição razoável, sendo assim uma boa primeira escolha ao atacar um problema de classificação. A principal vantagens das **redes neurais** é sua capacidade de fazer predições robustas: isso se dá devido ao grande número de hiperparametros (numero de nós na camada oculta, numero de camadas ocultas, algoritmo de otimização, função de otimização).

---

4. No contexto de machine learning e de acordo com o que discutimos, qual é a diferença entre validação e teste? Explique.

R.: O conjunto de validação é usado para mensurar uma estimativa do Erro out of sample, com isso podemos comparar diversos modelos candidatos. O conjunto de teste, por sua vez, representa os dados out of sample e é utilizado para de fato mensurar a performance do modelo (escolhido no processo de validação) em um conjunto grande de dados não vistos antes.

5. O que você entende por overfitting? Como podemos detectar overfitting e como podemos combatê-lo ?

   R.: Overfitting é um comportamento do modelo quando o aprendizado no conjunto de treino é tão intenso que o modelo **memoriza inclusive os ruídos nos dados**. Podemos analisar o overfitting pela capacidade de generalização do modelo: se a acurácia (ou qualquer outra métrica de avaliação do modelo) for muito alta no conjunto de treino e baixa no conjunto de validação ou teste, significa que o modelo não generalizou bem e aprendeu inclusive os ruídos no conjunto de treino. Existem duas principais estratégias para combater overfitting:

   1. Regularização
   2. Validação

Parte 2

6. Levando em consideração a lista acima, avalie o seu nível de aprendizado ao longo do semestre, respondendo as seguintes perguntas.
   - Qual é a porcentagem de tópicos que você acredita ter compreendido bem ? [0% a 100%]

     90%

   - Destaque um tópico sobre o qual você considera que mais avançou o seu conhecimento e
     disserte brevemente sobre ele.

     Validação, Cross Validation e Seleção de modelos. Antes entendia o conceito, sabia a utilidade de um conjunto de validação, mas nunca tinha efetuado um treinamento de validação cruzada antes. Foi uma boa experiência com mãos na massa de selecionar um modelo baseado nos dados, com todo cuidado pra evitar data leakage. Vale uma menção também a teoria do conhecimento. Não sabia nada disso antes e recebi várias justificativas do porquê os modelos de fato aprendem

   - Destaque dois tópicos sobre os quais você considera que menos avançou o seu conheci-
     mento e explique o porquê.

     SVMs e comparação entre aprendizado supervisionado e não supervisionado. Acho que não aproveitei tão bem essa parte do curso por estar no fim do semestre e acabou ficando corrido. Mas vou tentar revisitar no futuro, são dois temas bem importantes.

   - De forma geral, considerando as respostas aos items acima, qual nota entre 0 e 10 você daria para o seu grau de aproveitamento?

     8/10

7. Levando em consideração a lista de tópicos acima, avalie a sua dedicação ao longo da disciplina, respondendo às seguintes perguntas.

- Porcentagem de tópicos que você estudou (ou assistindo as aulas – online ou offline, ou vendo os vídeos da Caltech, ou lendo o livro-texto, ou buscando e estudando material extra) [0% a 100%].

  70%

- Quantos QT, Lista, ou EP deixou de fazer ou entregou com atraso? Se houve, quais são as justificativas dos atrasos ou da não-entrega?

  Deixei de entregar 1 QT. Não entreguei porque esqueci do prazo

- De forma geral, considerando as respostas aos items acima, se sua dedicação fosse tradu-
  zida para frequência (presença), qual seria a sua frequência ? [0% a 100%]

  80%

8. Avaliar é verificar se os objetivos foram alcançados. Nesta questão, levando em consideração suas expectativas iniciais, faça uma auto-avaliação sobre o seu desempenho na disciplina. Aborde sua facilidade com o conteúdo da disciplina, seu entendimento dos conceitos apresentados, sua assiduidade às aulas, sua participação em aulas e seu desempenho nas tarefas.
   Fique à vontade para adicionar outros comentários.

Acho que aproveitei bem a disciplina. Entrei nela com um certo know how de Machine Learning mas a disciplina entrou bem pra ensinar alguns formalismos que me faltaram. Me dediquei bem aos EPs, Listas e questionários e assisti a maioria das aulas da Caltech e algumas aulas gravadas. Sempre que necessário consultei o livro texto também.

Os EPs ajudaram imensamente. Ver o código funcionando de acordo com a teoria foi bem satisfatório. Achei que os QTs não foram bem estruturados. Acredito que, ao inves dos QTs, uma presença maior de listas seria bem interessante, já que elas me fizeram parar e pensar bastante sobre fundamentos teóricos ou revisitar a literatura.

No geral achei a disciplina muito bem estruturada. Meu único comentário foi termos ficado muito tempo falando de formalismo em ML de forma que muitos temas legais no fim do semestre (CNN, BOW, Aprendizado não supervisionado) ficaram bem apertados. No entanto, essa discussão sobre teoria do aprendizado é muito importante e, como disse antes é o que mais me agregou. Acho que esse comentário na verdade é mais um desejo de existir uma matéria de Machine Learning 2 no currículo =)