

Aprendizado de Máquina

Aula 11

<http://www.ic.uff.br/~bianca/aa/>

Tópicos

1. Introdução – Cap. 1 (16/03)
2. Classificação Indutiva – Cap. 2 (23/03)
3. Árvores de Decisão – Cap. 3 (30/03)
4. Ensembles - Artigo (13/04)
5. Avaliação Experimental – Cap. 5 (20/04)
6. Aprendizado de Regras – Cap. 10 (27/04)
7. Redes Neurais – Cap. 4 (04/05)
8. Teoria do Aprendizado – Cap. 7 (11/05)
9. Máquinas de Vetor de Suporte – Artigo (18/05)
10. Aprendizado Bayesiano – Cap. 6 e novo cap. online (25/05)
11. **Aprendizado Baseado em Instâncias – Cap. 8 (01/05)**
12. Classificação de Textos – Artigo (08/06)
13. Aprendizado Não-Supervisionado – Artigo (15/06)

Aula 11 - 01/06/2010

2

Aprendizado baseado em instâncias

- Ao contrário das outras abordagens, não ocorre a construção de um modelo de classificação explícito.
 - Novos exemplos são classificados com base na comparação direta e similaridade aos exemplos de treinamento.
- Treinamento pode ser fácil, apenas memorizar exemplos.
- Teste pode ser caro pois requer comparação com todos os exemplos de treinamento.
- Também conhecido como:
 - Aprendizado baseado em casos
 - Aprendizado baseado em exemplos
 - Vizinho mais próximo (“Nearest Neighbor” = kNN)
 - Aprendizado baseado em memorização
 - Aprendizado “lazy”

Aula 11 - 01/06/2010

3

Medidas de Similaridade/Distância

- Métodos baseados em instância precisam de uma função para determinar a similaridade ou distância entre duas instâncias.
- Para atributos contínuos, distância euclidiana é a mais utilizada:

$$d(x_i, x_j) = \sqrt{\sum_{p=1}^n (a_p(x_i) - a_p(x_j))^2}$$

onde $a_p(x)$ é o valor do p -ésimo atributo da instância x .
- Para atributos discretos, a distância é 0 se eles tem o mesmo valor e 1 se eles são diferentes.
- Para compensar as diferenças de unidade entre os atributos contínuos, temos que normalizá-los para ficar no intervalo $[0,1]$.

Aula 11 - 01/06/2010

4

Outras Medidas de Distância

- Distância de Mahalanobis
 - Métrica que faz uma normalização em relação à variância.
- Similaridade de Cosseno
 - Cosseno do ângulo entre os vetores.
 - Usado para classificação de textos e outros dados de alta dimensão.
- Correlação de Pearson
 - Coeficiente de correlação usado em estatística.
 - Muito usado em bioinformática.
- Distância de edição
 - Usado para medir distância entre strings.
 - Usado em classificação de textos e bioinformática.

Aula 11 - 01/06/2010

5

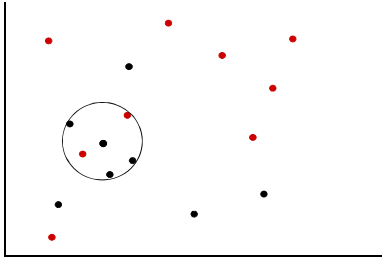
K-Vizinhos Mais Próximos

- Calcular a distância entre o exemplo de teste e cada exemplo de treinamento.
- Escolher os k vizinhos mais próximos e classificar a instância de teste com a classe mais frequente entre os vizinhos mais próximos.
- Usar mais de um vizinho reduz a vulnerabilidade a ruídos.
 - Porém um número muito grande de vizinhos pode incluir “vizinhos” distantes.
 - Melhor valor de k depende dos dados.
 - Normalmente usamos k ímpar pra evitar empates.

Aula 11 - 01/06/2010

6

Exemplo: 5-NN

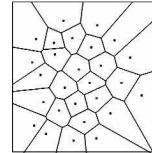


Aula 11 - 01/06/2010

7

Função de Classificação Implícita

- Embora não seja necessário calculá-la, a regra de classificação implicitamente usada é baseada em regiões do espaço de atributos, ao redor de cada exemplo de treinamento.
- Para 1-NN com distância euclidiana, o **diagrama de Voronoi** mostra como o espaço é dividido:



Aula 11 - 01/06/2010

8

Indexação Eficiente

- Uma busca linear para encontrar os vizinhos mais próximos não é eficiente para bases grandes.
- Estruturas de indexação podem ser usadas para acelerar a busca.
- Para distância euclidiana, uma estrutura chamada **kd-tree** pode ser usada para encontrar os vizinhos mais próximos em tempo $O(\log n)$.
 - Nós testam valores dos atributos e folhas terminam nos vizinhos mais próximos.
- Outras estruturas podem ser utilizadas para outras medidas de distância
 - Índice invertido para texto.

Aula 11 - 01/06/2010

9

Variações do k-NN

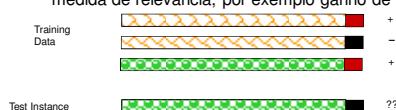
- Pode ser usado para estimar o valor de uma função **real** (regressão) se tirarmos a **média** dos valores dos k vizinhos mais próximos.
- Todos os exemplos de treinamento podem ser usados para classificar uma instância se cada um votar com **peso** proporcional ao inverso da sua distância ao exemplo de teste.

Aula 11 - 01/06/2010

10

Relevância dos Atributos

- Medidas de distância dão peso igual a todos os atributos mesmo que eles não sejam úteis para classificação.
 - Problemático quando muitos atributos são irrelevantes.
- Métodos baseados em instância favorecem a similaridade global e não a simplicidade do conceito.
 - Para simplicidade, necessário fazer uma seleção de atributos prévia.
 - Podemos dar peso aos atributos de acordo com uma medida de relevância, por exemplo ganho de informação.

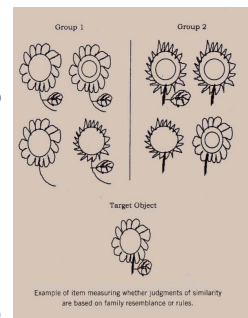


Aula 11 - 01/06/2010

11

Regras e Instâncias no Aprendizado Humano

- Experimentos psicológicos mostram que culturas diferentes têm diferentes vieses no aprendizado.
- “Ocidentais” favorecem regras simples (cabo reto) e classificam o objeto alvo no grupo 2.
- “Orientais” favorecem similaridade global e classificam o objeto alvo no grupo 1.



Aula 11 - 01/06/2010

12

Outros comentários

- Podemos reduzir a base de dados a um conjunto pequeno de exemplos representativos.
 - Análogos aos vetores de suporte no SVM.
 - K-NN condensado
- Podemos criar classificadores “híbridos”.
 - Usando gaussianas em redes neurais e SVMs estamos levando em consideração a vizinhança quando criamos os classificadores.
- Pode ser facilmente adaptado pra outros tipos de dados além de vetores (como grafos e strings).
 - O cálculo da similaridade também será complexo, dependendo do tipo de objeto.
- Pode ser usado pra outras tarefas de IA além de classificação.
 - Planejamento baseado em casos.
 - Raciocínio baseado em casos.

Aula 11 - 01/06/2010

13

k-NN Condensado

- Algoritmo incremental: Adicionar instância se necessário

```

 $\mathcal{Z} \leftarrow \emptyset$ 
Repeat
  For all  $\mathbf{x} \in \mathcal{X}$  (in random order)
    Find  $\mathbf{x}' \in \mathcal{Z}$  s.t.  $\|\mathbf{x} - \mathbf{x}'\| = \min_{\mathbf{x}^j \in \mathcal{Z}} \|\mathbf{x} - \mathbf{x}^j\|$ 
    If  $\text{class}(\mathbf{x}) \neq \text{class}(\mathbf{x}')$  add  $\mathbf{x}$  to  $\mathcal{Z}$ 
Until  $\mathcal{Z}$  does not change
  
```

Aula 11 - 01/06/2010

14