Aprendizado de Máquina

Aula 4

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. Teoria do Aprendizado – Cap. 7 (27/04)
7. Aprendizado de Regras – Cap. 10 (04/05)
8. Redes Neurais – Cap. 4 (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)

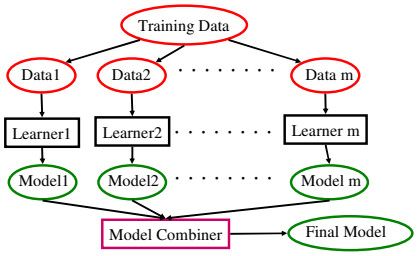
Aprendizado por agrupamento

(Russell & Norvig – 18.4, Hastie & Tibshirani – 8.7)

- Aprendizado por agrupamento = *Ensemble learning* (*ensemble* = junto em francês).
- Os métodos tradicionais de aprendizado, selecionam uma única hipótese, escolhida de um espaço de hipóteses.
  - Ex.: Escolhe-se a menor árvore de decisão que seja consistente com os dados.
- Os métodos de aprendizado por agrupamento selecionam uma coleção de hipóteses e combinam as suas previsões.
  - Ex.: Gera-se 100 árvores de decisão diferentes que votam na melhor classificação.

Aprendizado de Agrupamentos

- Aprende-se múltiplas definições de um conceito usando dados de treinamento diferentes OU algoritmos diferentes.
- Combinam-se as decisões, por exemplo, usando votação.



Motivação para o Agrupamento

- Quando combinamos decisões ***independentes*** e ***diferentes*** sendo que cada uma é mais precisa do que um "chute", erros aleatórios se cancelam e decisões corretas são reforçadas.
- Agrupamentos humanos
  - "Quem quer ser um milionário?": Ajuda do amigo vs. voto da plateia.

Motivação para o Agrupamento

- Exemplo:
  - Considere um conjunto de k=5 hipóteses.
  - Suponha que combinamos suas previsões usando maioria simples.
  - Para o conjunto classificar o exemplo de modo incorreto, 3 das 5 hipóteses devem estar incorretas.
- Supondo que cada hipótese h_i tem uma taxa de erro p e que os erros sejam independentes, com k hipóteses o erro será $p^{\lfloor k/2 \rfloor + 1}$.
  - Se p=0.1 e k=5, a nova taxa de erro será 0.001.
- Na prática, os erros não serão totalmente independentes, pois as hipóteses serão obtidas usando os mesmos dados.
  - Mesmo assim, se as hipóteses forem ao menos um pouco diferentes, reduzindo a correlação dos erros, o agrupamento diminui a taxa de erros.

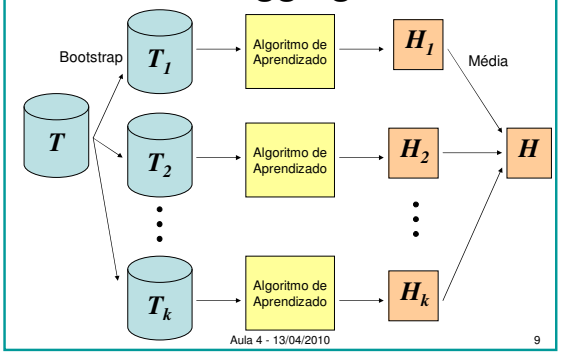
Bagging

- Bagging = **B**ootstrap **Agreggat**i**ng**
- Um dos métodos de agrupamento mais simples.
  - Criado por Leo Breiman em 1996.
  - Baseia-se na criação de *bootstraps* = amostras diferentes da base de dados que são usadas para aprender hipóteses diferentes.
  - A previsão final para um exemplo de teste é a média da previsão de cada hipótese.
  - Pode ser usado tanto para classificação quanto para regressão.
    - No caso de classificação, média é equivalente a maioria.
  - Funciona muito bem na prática, para alguns algoritmos de aprendizado.

Bootstrap

- Dada uma base de treinamento T com n exemplos, geramos k novas bases T_k também com n exemplos cada através de amostragem com reposição.
  - Para gerar cada T_k, sorteamos um número de 1 a n, n vezes, e a cada vez adicionamos o exemplo $T[n]$ a T_k.
  - Como resultado, cada base T_k terá aproximadamente 66% dos exemplos de T, com duplicatas.

Bagging



Quando Bagging funciona?

- Um fator crítico para que o Bagging funcione é a instabilidade do algoritmo de aprendizado.
- Se pequenas mudanças na base de treinamento T provocarem grandes mudanças na hipótese H gerada, dizemos que o algoritmo de aprendizado é instável.
  - Ex: árvores de decisão e redes neurais.
- Caso contrário, ele é estável.
  - Ex.: k-NN, regressão linear.
- Bagging funciona melhor quanto maior for a instabilidade do algoritmo.

Desvantagens de Bagging

- Perde-se na interpretação.
  - Em vez de uma única hipótese, temos uma combinação, que é mais difícil de ser entendida.
  - Por exemplo, uma combinação de árvores de decisão não pode ser transformada numa única árvore.
- Tem um custo computacional adicional.
- Só funciona quando o algoritmo é instável.

Observações sobre Bagging

- Conforme adicionamos mais hipóteses, mais devagar cai a taxa de erro, até que ela se estabiliza.
- Tirar a média das estimativas de probabilidade é melhor do que tirar a média das previsões.
- Em árvores de decisão, a poda aumenta a estabilidade.
  - Logo bagging funciona melhor sem a poda (*pruning).

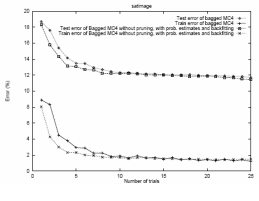


Figura de Bauer e Kohavi, 1999.

Alguns resultados (Maclin, 1997)

Table 2: Test set error rates for the data sets using (1) a single neural network classifier; (2) an ensemble where each individual network is trained using the original training set and thus only differs from the other networks in the ensemble by its random initial weights; (3) an ensemble where the networks are trained using randomly resampled training sets (Bagging); an ensemble where the networks are trained using weighted resampled training sets (Boosting) where the resampling is based on the (4) Arcing method and (5) Ada method; (6) a single decision tree classifier; (7) a Bagging ensemble of decision trees; and (8) a Boosting ensemble of decision trees.

	Neural Network					C4.5		
				Boosting				Boosting
Dataset	Standard	Simple	Bagging	Arcing	Ada	Standard	Bagging	Ada
breast-cancer-w	3.3	3.4	3.3	3.2	3.9	5.0	3.3	3.1
credit-a	14.8	14.0	14.1	14.8	16.2	14.9	12.1	12.6
credit-g	28.3	24.4	24.3	24.8	26.4	29.6	22.8	22.9
diabetes	23.6	22.8	23.2	23.6	22.8	28.3	21.9	22.3
glass	38.5	35.5	33.7	31.5	33.2	30.9	28.4	30.5
heart-cleveland	18.2	17.3	16.7	18.9	19.1	24.3	18.1	17.4
hepatitis	19.9	19.6	18.1	18.9	18.1	21.6	16.5	13.8
house-votes-84	5.0	4.9	4.3	4.7	5.1	3.5	3.6	4.4
hypo	6.4	6.2	6.2	6.4	6.2	0.5	0.4	0.4
ionosphere	10.1	8.0	7.6	7.0	8.0	8.1	6.0	6.0
iris	4.3	4.0	4.3	2.9	3.3	6.0	4.6	5.6
kr-vs-kp	2.3	0.9	0.9	0.6	0.4	0.6	0.5	0.3
labor	5.3	4.2	4.9	3.2	5.0	15.1	13.3	13.2
letter	18.0	12.8	12.5	6.2	4.6	14.0	10.6	6.7
promoters-936	5.0	4.8	4.5	4.7	4.7	12.8	9.5	6.3
ribosome-bind	9.5	8.5	8.4	8.4	8.5	11.2	9.3	9.1
satellite	12.9	11.1	11.0	10.3	10.3	13.8	10.8	10.4
segmentation	6.7	5.6	5.3	3.7	3.7	3.7	2.8	2.3
sick	6.0	5.7	5.8	5.4	4.8	1.3	1.0	0.9
sonar	16.9	16.7	16.5	15.9	12.5	29.0	21.6	19.7
soybean	9.0	6.4	6.8	6.5	6.3	8.0	8.0	7.9
splice	4.7	4.0	3.9	4.0	4.3	5.9	5.7	6.3
vehicle	24.5	21.1	21.7	19.3	19.5	29.4	26.1	24.8

Boosting

- A idéia do algoritmo de agrupamento Boosting (também chamado de aceleração) é construir hipóteses sucessivas, de tal modo que exemplos classificados incorretamente por hipóteses anteriores sejam melhor classificados por hipóteses seguintes.
- Desenvolvido em 1996 por Freund e Schapire.

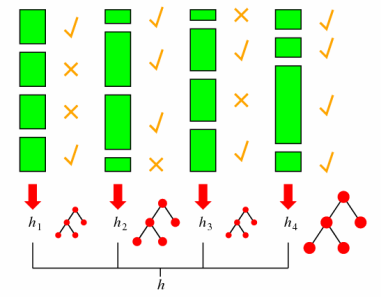
Requerimentos para o boosting

- Serve apenas para classificação.
- Supõe que o algoritmo de aprendizado utilizado trabalhe com conjuntos de treinamento ponderados.
  - Cada exemplo tem um peso associado $w_j \geq 0$.
  - Quanto mais alto o peso de um exemplo, mais alta deve ser importância dada a ele pelo algoritmo (parecido com aprendizado sensível a custos).
    - Algoritmos com Naive Bayes e árvores de decisão podem ser facilmente modificados para incorporar pesos.
    - Caso não seja possível incorporá-los pode-se usar amostragem de acordo com os pesos.

A idéia geral de boosting

- Boosting começa com $w_j = 1$ para todos os exemplos da base.
- A partir desse conjunto, ele gera a primeira hipótese, H_1.
  - Essa hipótese classificará alguns dos exemplos de treinamento corretamente e outros incorretamente.
- Para que a próxima hipótese classifique melhor os exemplos incorretamente classificados, aumentamos os seus pesos, diminuindo o peso daqueles corretamente classificados.
- A partir desse novo conjunto de treinamento com pesos, geramos a hipótese H_2.
- O processo continua até que sejam geradas k hipóteses.
- Existem muitas variações de boosting dependendo de como os pesos são atribuídos e de como as hipóteses são combinadas.
  - A mais conhecida é o AdaBoost.

A idéia geral de boosting



AdaBoost

function ADABOOST(*examples*, L, M) **returns** a weighted-majority hypothesis
  inputs: *examples*, set of N labelled examples $(x_1, y_1), \ldots, (x_N, y_N)$
    L, a learning algorithm
    M, the number of hypotheses in the ensemble
  local variables: **w**, a vector of N example weights, initially $1/N$
    h, a vector of M hypotheses
    z, a vector of M hypothesis weights

  for $m = 1$ **to** M **do**
    $\mathbf{h}[m] \leftarrow L(examples, \mathbf{w})$
    $error \leftarrow 0$
    for $j = 1$ **to** N **do**
      if $\mathbf{h}[m](x_j) \neq y_j$ **then** $error \leftarrow error + \mathbf{w}[j]$
    for $j = 1$ **to** N **do**
      if $\mathbf{h}[m](x_j) = y_j$ **then** $\mathbf{w}[j] \leftarrow \mathbf{w}[j] \cdot error/(1 - error)$
    $\mathbf{w} \leftarrow$ NORMALIZE(**w**)
    $\mathbf{z}[m] \leftarrow \log(1 - error)/error$
  return WEIGHTED-MAJORITY(**h**, **z**)

Figure 18.12

Propriedades do AdaBoost

- Se o algoritmo de aprendizado utilizado garante aprendizagem fraca, então o AdaBoost retornará uma hipótese que classifica perfeitamente os dados de treinamento.
- Aprendizagem fraca: melhor do que classificar os dados aleatoriamente.
- Esse resultado é válido independentemente do espaço de hipóteses original e da complexidade da função que está sendo aprendida.

Desvantagens do AdaBoost

- Ao contrário de bagging, há o risco de super-ajuste (over-fitting).
  - Principalmente em casos onde há muito ruído na base de dados.
  - Caso haja erros de classificação nos dados de treinamento, boosting vai colocar muito peso nesses dados.

Alguns resultados (Maclin, 1997)

Table 2: Test set error rates for the data sets using (1) a single neural network classifier; (2) an ensemble where each individual network is trained using the original training set and thus only differs from the other networks in the ensemble by its random initial weights; (3) an ensemble where the networks are trained using randomly resampled training sets (Bagging); an ensemble where the networks are trained using weighted resampled training sets (Boosting) where the resampling is based on the (4) Arcing method and (5) Ada method; (6) a single decision tree classifier; (7) a Bagging ensemble of decision trees; and (8) a Boosting ensemble of decision trees.

	Neural Network					C4.5		
				Boosting				Boosting
Dataset	Standard	Simple	Bagging	Arcing	Ada	Standard	Bagging	Ada
breast-cancer-w	3.3	3.4	3.3	3.2	3.9	5.0	3.3	3.1
credit-a	14.8	14.0	14.1	14.8	16.2	14.9	12.1	12.6
credit-g	28.3	24.4	24.3	24.8	26.4	29.6	22.8	22.9
diabetes	23.6	22.8	23.2	23.6	22.8	28.3	21.9	22.3
glass	38.5	35.5	33.7	31.5	33.2	30.9	28.4	30.5
heart-cleveland	18.2	17.3	16.7	18.9	19.1	24.3	18.1	17.4
hepatitis	19.9	19.6	18.1	18.9	18.1	21.6	16.5	13.8
house-votes-84	5.0	4.9	4.3	4.7	5.1	3.5	3.6	4.4
hypo	6.4	6.2	6.2	6.4	6.2	0.5	0.4	0.4
ionosphere	10.1	8.0	7.6	7.0	8.0	8.1	6.0	6.0
iris	4.3	4.0	4.3	2.9	3.3	6.0	4.6	5.6
kr-vs-kp	2.3	0.9	0.9	0.6	0.4	0.6	0.5	0.3
labor	5.3	4.2	4.9	3.2	5.0	15.1	13.3	13.2
letter	18.0	12.8	12.5	6.2	4.6	14.0	10.6	6.7
promoters-936	5.0	4.8	4.5	4.7	4.7	12.8	9.5	6.3
ribosome-bind	9.5	8.5	8.4	8.4	8.5	11.2	9.3	9.1
satellite	12.9	11.1	11.0	10.3	10.3	13.8	10.8	10.4
segmentation	6.7	5.6	5.3	3.7	3.7	3.7	2.8	2.3
sick	6.0	5.7	5.8	5.4	4.8	1.3	1.0	0.9
sonar	16.9	16.7	16.5	15.9	12.5	29.0	21.6	19.7
soybean	9.0	6.4	6.8	6.5	6.3	8.0	8.0	7.9
splice	4.7	4.0	3.9	4.0	4.3	5.9	5.7	6.3
vehicle	24.5	21.1	21.7	19.3	19.5	29.4	26.1	24.8