

Análise de Imagens

Aula 19: Árvores de Decisão

Prof. Alexandre Xavier Falcão

afalcao@ic.unicamp.br

IC - UNICAMP

Roteiro da Aula

- Árvore de decisão.
- Crescimento da árvore (aprendizado).
- Seleção de característica para um dado nó.
- Critérios de parada.
- Métodos de poda.

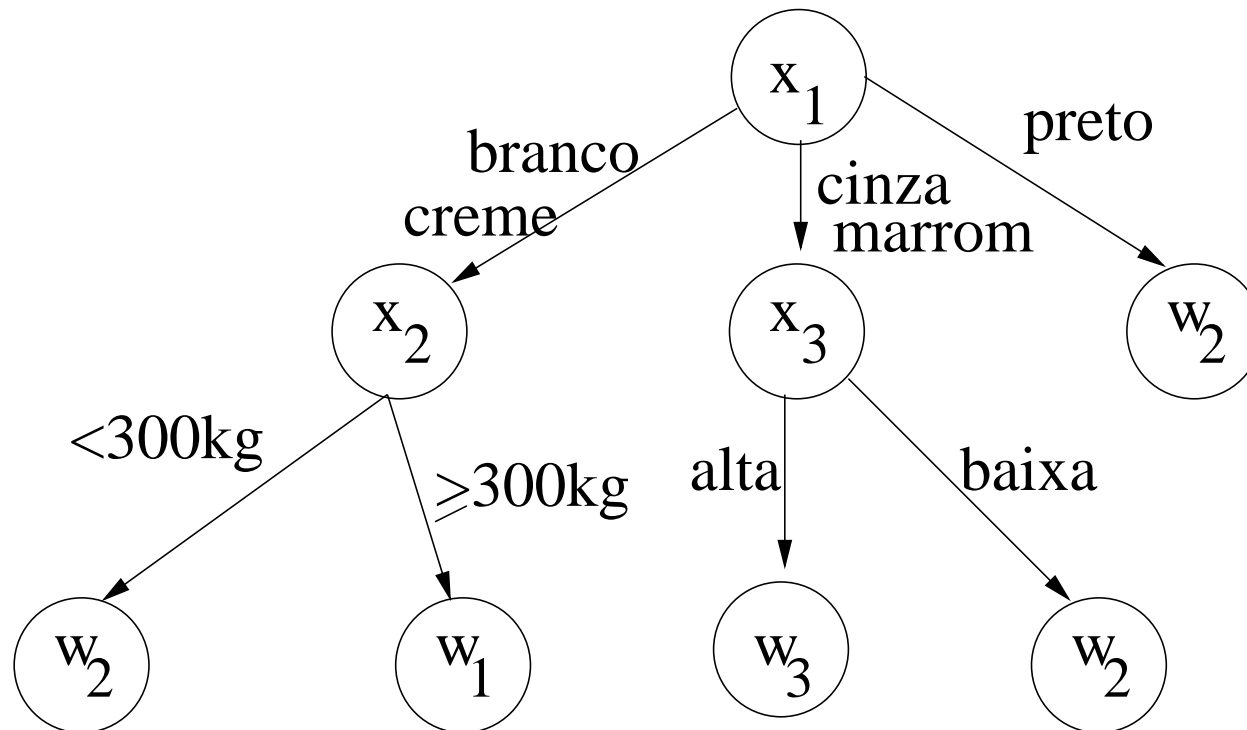
Ver livro da Kuncheva.

Árvore de decisão

Seja $\{x_1, x_2, \dots, x_n\}$ um conjunto de características, as quais podem ser quantitativas ou qualitativas. Considere como processo de decisão uma seqüência finita de passos (caminho em uma árvore), que inicia em um nó raíz e leva uma amostra \mathbf{x} a uma dada classe w_i (nó folha). A cada passo, uma das n características é selecionada e dependendo do seu valor na amostra, o passo seguinte é um nó filho de possíveis ramos da árvore para o nó corrente. Na árvore, uma classe pode ser representada por mais de um nó folha, e as folhas podem ou não estar no mesmo nível.

Árvore de decisão

Considere, por exemplo, uma classificação de ursos de acordo com três características; $x_1 = \text{cor}$, $x_2 = \text{peso}$, e $x_3 = \text{possibilidade de atacar humanos}$; e três classes; $w_1 = \text{ursus maritimus (polar)}$, $w_2 = \text{ursus americanus}$, e $w_3 = \text{ursus arctos (pardo)}$. A árvore correspondente é



Árvore de decisão

Se não ocorrer características de mesmo valor em classes distintas, o erro da árvore é zero. Este fato coloca as árvores de decisão na categoria de classificadores instáveis. Pequenas alterações nos dados de treinamento podem levar a árvores bem diferentes, o que favorece a combinação de classificadores usando árvores de decisão. A mesma observação é válida para redes neurais. Para automatizar sua construção, o modelo mais adequado é o de uma árvore binária. A cada passo, uma característica é escolhida e os dados de treinamento são divididos em dois ramos.

Árvore de decisão

O processo se repete e quando um nó recebe a maioria expressiva de amostras de uma dada classe, este nó é selecionado como folha dessa classe. A maioria expressiva define um grau de impureza baixo da distribuição probabilística das classes no nó. O nó é definido como folha com base nesta medida. A seleção de uma característica em um dado nó também se baseia no grau de impureza. A característica selecionada deve ser aquela que mais reduz o grau de impureza dos nós filhos com relação ao grau de impureza do nó corrente.

Grau de impureza

Seja P_i a probabilidade associada à classe w_i , $i = 1, 2, \dots, c$, em um dado nó t , a qual é medida pela proporção de amostras de treinamento que são da classe w_i entre as amostras que chegam ao nó t . As fórmulas abaixo medem o grau de impureza $i(t)$ no nó t .

$$i_1(t) = - \sum_{i=1}^c P_i \log P_i$$

$$i_2(t) = 1 - \sum_{i=1}^c P_i^2$$

$$i_3(t) = 1 - \max_{i=1}^c \{P_i\}$$

Grau de impureza

A primeira equação mede a entropia em t . O valor de $i_1(t)$ é mínimo em 0 quando só existem amostras de uma dada classe. Se a distribuição de amostras por classe for uniforme, $P_i = \frac{1}{c}$, $i = 1, 2, \dots, c$, $i_1(t)$ terá valor máximo $\log(c)$. A segunda equação é a impureza de Gini. Os valores mínimo e máximo de $i_2(t)$ são 0 e $\frac{c-1}{c}$ nas situações de amostras de uma única classe e distribuição uniforme, respectivamente. A terceira equação representa o erro esperado de classificação se o nó fosse definido como folha e a classe escolhida fosse a de maior P_i .

Seleção de característica

Para dividir t em dois filhos t_0 e t_1 com base em uma dada característica x_j , $1 \leq j \leq n$, o ganho desta divisão é a queda de impureza medida por

$$\Delta i(t) = i(t) - P_{t_0}i(t_0) - P_{t_1}i(t_1)$$

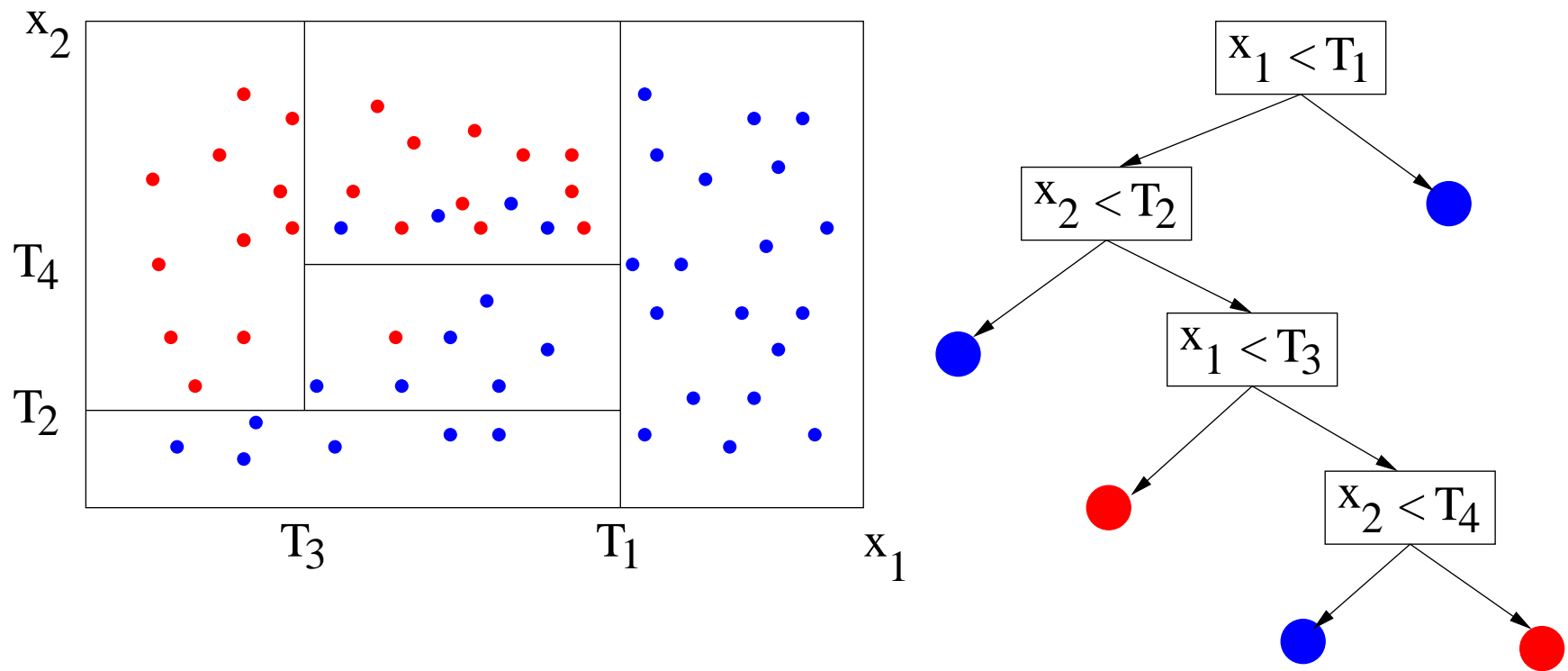
onde P_{t_k} , $k = 1, 2$, é a probabilidade de escolher t_k estando em t . Escolhemos x_j que produz maior $\Delta i(t)$. Se x_j for contínua, temos ainda que verificar qual é o limiar ótimo para dividir as amostras em t . Por exemplo, $x_j < T$ e $x_j \geq T$ divide as amostras em dois nós, mas temos que achar T onde a redução de impureza é máxima.

Seleção de característica

O valor de T é encontrado ordenando-se x_j para todas amostras em t . Ao variar T para os valores da seqüência ordenada, estamos dividindo as amostras que estão acima e abaixo de T . Estas divisões geram os possíveis nós filhos, e buscamos o valor de T onde $\Delta i(t)$ é máximo. O procedimento deve ser repetido para cada característica $x_j, j = 1, 2, \dots, n$, encontrando assim o valor ótimo de T para cada característica e a característica com a maior redução de impureza.

Exemplo

O treinamento finaliza com a construção da árvore. Um exemplo é mostrado na figura abaixo.



Critérios de parada

A construção da árvore poderia parar apenas quando não houvessem mais nós impuros. Se os dados de treinamento não contivessem objetos com características de mesmo valor em classes distintas, a árvore seria perfeita. Contudo, o resultado perfeito poderia significar supertreinamento do classificador e péssimo desempenho nos testes. Portanto, critérios de parada são adotados para evitar nós puros como folha. Ao mesmo tempo devemos evitar subtreinamento com paradas antes da hora.

Critérios de parada

Algumas sugestões incluem o uso de um conjunto Z_2 de validação tal que paramos de dividir os nós quando o erro da classificação usando Z_2 aumenta. Outra idéia estabelece um limiar β tal que a parada ocorre para um nó t quando a maior redução possível de impureza em t for menor que β . O problema é escolher β . Um variante um pouco melhor é estabelecer um limiar no número mínimo de amostras por nó da árvore. Este limiar pode ser um percentual de $|Z_1|$. Entre todas as idéias, a mais interessante parece ser a baseada em teste de hipótese.

Teste de hipótese

Considerando que em um dado nó t nós temos n amostras, sendo n_i amostras de cada classe w_i , e que a divisão de t em dois nós l e r divide estas amostras em n^l e n^r amostras, sendo n^l_i e n^r_i amostras da classe w_i para cada nó, respectivamente. Nosso problema é saber se esta divisão tem significância estatística. Ou seja, se as distribuições das amostras de cada classe nos nós filhos forem equivalentes à distribuição no nó pai, a divisão não ocorre. As medidas

$$\chi_l^2 = \sum_{i=1}^c \frac{(n \times n^l_i - n^l \times n_i)^2}{n \times n^l \times n_i}$$

$$\chi_r^2 = \sum_{i=1}^c \frac{(n \times n^r_i - n^r \times n_i)^2}{n \times n^r \times n_i}$$

Teste de hipótese

são obtidas entre os filhos e o pai. O valor médio $\chi^2 = \frac{\chi_l^2 + \chi_r^2}{2}$ é comparado com o valor tabulado para uma dada significância estatística e grau 1 de liberdade. Se χ^2 for maior que o tabulado, a divisão se justifica. Os critérios de parada são denominados métodos de pré-poda. Algumas vezes, porém, é mais vantagem deixar a árvore crescer completamente e depois realizar a poda. A pós-poda tem como objetivo balancear o aumento do erro de treinamento com a redução do tamanho da árvore, evitando supertreinamento. Estes métodos são chamados métodos de poda.

Poda por erro reduzido

Seja Z_p um conjunto adicional, chamado conjunto de poda, o qual não possui amostras em Z_1 . Considere os nós internos de cada caminho que sai de uma folha para a raiz da árvore. Para um nó interno t , avalie o erro de classificação sobre Z_p caso t fosse um nó folha cuja classe fosse definida pela maioria das amostras de Z_1 em t . Se este erro for menor que o erro da árvore completa sobre Z_p , então t vira nó folha. Caso contrário, a subárvore é mantida. O procedimento se repete para os demais nós internos em direção à raiz. A árvore obtida possui erro mínimo com relação a Z_p .

Poda por erro pessimista

Iniciando na raíz, cada nó interno é avaliado em direção às folhas. Seja t um nó interno; n o número de amostras de Z_1 em t ; $e(t)$ o erro sobre Z_1 se t for trocado por uma folha, usando a maioria para definir a classe; T_t a subárvore de t ; \mathcal{L}_t o conjunto de folhas de T_t ; e $e'(T_t) = \sum_{l \in \mathcal{L}_t} e(l) + \frac{|\mathcal{L}_t|}{2}$. O nó t é trocado por uma folha se $e(t) \leq e'(T_t) + \sqrt{\frac{e'(T_t)[n - e'(T_t)]}{n}} - \frac{1}{2}$. O método busca um balanço entre erro e número de folhas, o qual não tem uma justificativa plausível. O livro discute outros métodos, mas eles dependem de parâmetros *ad hoc*. O erro reduzido parece ser o mais interessante de todos.