ACH2016 - Inteligência Artificial
Aula 09 - Aprendizado Supervisionado além da Classificação Binária

Valdinei Freire da Silva

valdinei.freire@usp.br - Bloco A1 100-O

Russell e Norvig, Capítulo 18

Tarefa de Aprendizado Supervisionado

Dado um conjunto de treinamento com N exemplos de pares entrada-saída

$$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N),$$

onde cada y_i foi gerado por uma função f desconhecida, isto é, $y_i = f(x_i)$.

Descubra uma função h que aproxima a verdadeira função f.

x é a entrada e y é a saída.

x e y pode ser qualquer valor, números ou categorias, x usualmente é um vetor de valores (atributos).

Melhor Hipótese

Genericamente pode-se pensar em uma função de perda:

$$L(h, \mathcal{E})$$

que avalia a qualidade da hipótese h aplicada na população $\mathcal{E}$.

A hipótese ótima é dada por:

$$h^* = \arg\min_{h \in \mathcal{H}} L(h, \mathcal{E}).$$

Empiricamente, para um conjunto de exemplos E, temos:

$$\hat{h}^* = \arg\min_{h \in \mathcal{H}} L(h, E).$$

Exemplo de função de perda (*Loss Function*):

- Acurácia: taxa de exemplos que são classificados corretamente.
- Verosimilhança: apenas para hipóteses probabilísticas.

Algoritmos Estudados

- Árvores de Decisão
- Classificador Linear
- Regressão Logística
- Redes Neurais
- k-Vizinhos mais Próximos

Problema 1

Entrada: notas nas disciplinas do primeiro ano.

Saída: semestres para se formar.

Filtro: apenas alunes que foram aprovades em todas as disciplinas e se formaram.

entrada x						saída y
FSI	RP I	Calc I	⋯	Calc II	MVGA	semestres
5.0	6.7	3.4	⋯	7.00	5.5	10
9.0	9.5	8.0	⋯	8.0	9.0	8
7.0	9.5	6.0	⋯	5.5	5.9	9
...	...	...	...	...	...	...

Problema 2

Entrada: notas nas disciplinas do primeiro ano.

Saída:

- não se formou (N)
- se formou e não iniciou mestrado em 2 anos (F)
- se formou e iniciou mestrado em 2 anos (M)

Filtro: apenas alunes que foram aprovades em todas as disciplinas.

entrada x						saída y
FSI	RP I	Calc I	⋯	Calc II	MVGA	resultado
5.0	6.7	3.4	⋯	7.00	5.5	N
9.0	9.5	8.0	⋯	8.0	9.0	F
7.0	9.5	6.0	⋯	5.5	5.9	M
...	...	...	...	...	...	...

Problema 3

Entrada: pares (alune, disciplina)

Saída: nota obtida pele alune na disciplina.

Filtro: qualquer alune ingressante desde de 2015.

	RP I	Calc I	⋯	Calc II	MVGA	IA
João	6.7	?	⋯	7.00	5.5	7.8
Maria	?	8.0	⋯	?	9.0	?
José	9.5	6.0	⋯	5.5	?	6.0
...	...	...	...	...	...	...

Estatística: Problema de Regressão

Em problemas de regressão, considera-se o seguinte modelo:

$$Y = f(\mathrm{X}) + \epsilon(\mathrm{X}),$$

onde X é um vetor de variáveis independentes observadas, Y é uma variável de resposta também observada, e $\epsilon(X)$ é uma variável aleatória com esperança 0.

O vetor aleatório X é chamado preditor, a variável aleatória Y é chamada de resposta. A esperança condicional de Y para um vetor dado x de X é chamada a **função regressão** de Y sobre X.

$$\mathrm{E}[Y|X = x] = f(x).$$

Estatística: Problema de Regressão

Um problema de regressão assume as seguintes suposições:

1. Ou os vetores $x_1, \ldots, x_n$ são conhecidos anteriormente no tempo ou eles são os valores observados de vetores aleatórios $X_1, \ldots, X_n$ sobre os quais é condicionada a distribuição conjunta de $Y_1, \ldots, Y_n$.
2. Para $i = 1, \ldots, n$, a distribuição condicional de Y_i dado os vetores $x_1, \ldots, x_n$ possuem a mesma variância (homocedasticidade).
3. Existe uma função $f : X \to \mathbb{R}$ tal que a esperança condicional de Y_i dados os vetores $x_1, \ldots, x_n$ tem a forma $E[Y_i | X_i = x_i] = f(x_i)$ para todo $i = 1, \ldots, n$.
4. As variáveis aleatórias $Y_1, \ldots, Y_n$ são independentes dada as observações $x_1, \ldots, x_n$.

Estatística: Regressão Linear Simples

Considere que o preditor $X \in \mathbb{R}$, e a função regressão $f(X) = \beta_0 + \beta_1 x$. Finalmente, considere que $\epsilon(X)$ independe de X e é uma variável com distribuição $N(0, \sigma)$. Um problema de regressão sob tais condições é chamado de regressão linear simples. A p.d.f. condicional de $Y_1, \ldots, Y_n$ é

$$f_n(\mathrm{y}|\mathrm{x}, \beta_0, \beta_1, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right].$$

Estimador de Máxima Verosimilhança:

$$\arg\max_{\beta_0,\beta_1} f_n(\mathrm{y}|\mathrm{x}, \beta_0, \beta_1, \sigma^2) = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Conclusão: $L(h, E) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h(x_i))^2$ - (M.S.E - Mean Square Error)

O M.L.E. do problema de regressão linear simples tem a seguinte solução:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

Para $i = 1, \ldots, n$, os valores observados $\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ são chamados de valores ajustados, e os valores observados de $e_i = y_i - \hat{y}_i$ são chamados de resíduos.

Estatística: Regressão Linear Múltipla

Considere que o preditor $X \in \mathbb{R}^p$, e a regressão $f(X) = \beta_0 X_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1}$. Finalmente, considere que $\epsilon(X)$ independente de X e é uma variável aleatória normal com média 0 e variância σ^2. Um problema de regressão sob tais condições é chamado de regressão múltipla e $f(X)$ é modelo linear múltiplo.

O M.L.E. do problema de regressão linear múltipla tem a seguinte solução:

$$\widehat{\beta} = (Z'Z)^{-1}Z'Y$$

$$\widehat{\sigma}^2 = \frac{1}{n}(Y - Z\beta)'(Y - Z\beta)$$

onde:

$$Z = \begin{bmatrix} x_{1,0} & \ldots & x_{1,p-1} \\ x_{2,0} & \ldots & x_{2,p-1} \\ \vdots & \ddots & \vdots \\ x_{n,0} & \ldots & x_{n,p-1} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \text{ e } \widehat{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}.$$

Funções Não-Lineares

Qual função melhor representa os dados?

Como realizar a regressão para polinômios?

- Cria o seguinte vetor de atributos: $\mathrm{x} = [1, x, x^2, x^3, \ldots, x^n]$
- Realiza regressão linear de Múltiplas Variáveis

Como realizar a regressão para redes Neurais?

- camada internas com função de ativação não-linear
- camada de saída linear
- busca por gradiente descendente

Filtros Colaborativos

Considere o seguinte problema de aprendizado de máquina:

- dado um conjunto de usuários $\mathcal{U}$ e um conjunto de itens $\mathcal{I}$
- considere o conjunto de exemplos $E = \{(U_1, I_1, y_i), \ldots, (U_n, I_n, y_n)\}$, onde $y_i \in \mathbb{R}$ é a avaliação atribuída ao item $I_i \in \mathcal{I}$ pelo usuário $U_i \in \mathcal{U}$
- **Objetivos:** construir a função $h : \mathcal{U} \times \mathcal{I} \to \mathbb{R}$ que modela a avaliação do usuário para qualquer item.
- **Dificuldade:** a descrição dos usuários e itens não possuem nenhuma estrutura, isto é, são enumerados.

Recomendação de Produtos

Descrição dos usuários e itens estruturada em atributos

- sistema de recomendação baseado em conteúdo
- aprende uma hipótese h_U para cada usuário
- técnicas típicas de aprendizado de máquina (redes neurais, árvore de decisão, K-NN, etc.)

Descrição dos usuários e itens enumerados

- sistema de recomendação baseado em filtros colaborativo
- aprende uma hipótese h conjunta para usuários e itens
- K-NN ou fatoração de matrizes

Filtros Colaborativos - K-NN

K-NN: considera uma noção de vizinhança, usualmente derivada de uma função de distância *Distance*
Dado um par (U, I) que se deseja a avaliação, a vizinhança são os K exemplos em E mais próximos de (U, I).

Opções:

- distância baseada em usuários
- distância baseada em itens
- combinação das duas avaliação

Filtros Colaborativos - K-NN

Distância baseada em Usuários

- considera-se apenas os usuários que avaliaram o item I, isto é,

$$\mathcal{U}_I = \{U' : \exists e_i = (U', I) \in E\}$$

- considera-se uma distância entre usuários
  - Distância Esparsa: considera apenas os itens avaliados em comum
  - Distância Densa: preenche os valores que faltam com avaliações médias
- escolhe os K vizinhos mais próximos no conjunto $\mathcal{U}_I$
- considera-se alguma média ponderada das avaliações dos K vizinhos para o item I.

Distância baseada em Itens (equivalente à distância baseada em usuários)

Fatoração de Matrizes

Considere que os usuários e itens são enumerados, então, cada usuário ou item pode ser representado por inteiro.

Considere uma matriz Y, na qual cada exemplo $e = (U, I, y)$ interpreta-se como sendo y o valor da célula da linha U e coluna I.

Note que a matriz Y é definida apenas parcialmente.

Considera-se o problema de encontrar matrizes $W \in \mathbb{R}^{|\mathcal{U}| \times k}$ e $F \in \mathbb{R}^{k \times |\mathcal{I}|}$ que minimizem:

$$||Y - WF||^p = \sum_{(U,I,y) \in E} |y - (WF)_{UI}|^p$$

Fatoração de Matrizes

Interpretação:

- WF aproxima (fatora) Y
- k é a quantidade de atributos latentes
- cada coluna de F representa os atributos encontrados para um item
- cada linha de W representa os pesos associados a cada atributo por um usuário
- note que os atributos de um item são compartilhados para todos os usuários

Fatoração de Matrizes Logística

Se as avaliações y são binárias, isto é, $y \in \{0, 1\}$, então pode-se utilizar regressão logística.

$$\min \sum_{(U,I,y)\in E} \left(y \times \log\left(\frac{1}{1 + e^{(WF)_{UI}}} \right) + (1 - y) \times \log\left(1 - \frac{1}{1 + e^{(WF)_{UI}}} \right) \right)$$

Exemplo: ENEM

Seja na regressão linear, ou na regressão logística, pode-se considerar o algoritmo de descida da encosta na direção do gradiente em W e F.

Classificação Multiclasse

Considere um conjunto de classes (categorias) $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$.

A função de classificação f leva uma entrada $x \in \mathcal{X}$ para uma classe em $\mathcal{C}$.

Um classificador baseado em várias regressões logísticas:

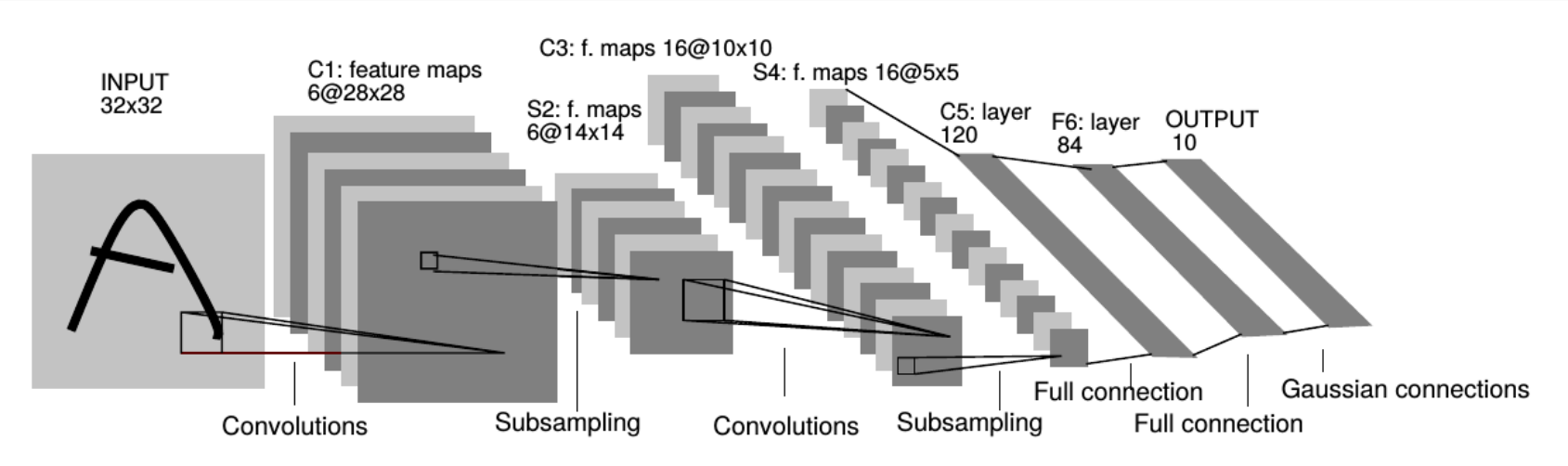
- para cada categoria C_i treina um modelo de **regressão logística** $h_i : \mathcal{X} \to \{0, 1\}$
- constrói um classificador $h(x) = \arg\max h_i(x)$

Um classificador baseado em várias funções de avaliação:

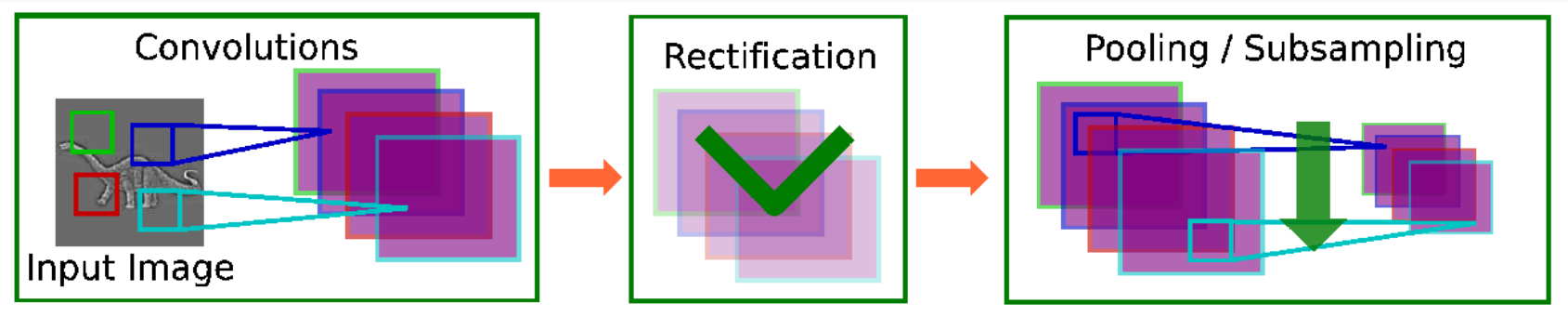
- para cada categoria C_i treina um modelo que obtém uma **função de avaliação** g_i
- constrói uma distribuição multinomial $h_i(x) = \frac{\exp(g_i(x))}{\sum_{j=1}^{k} \exp(g_j(x))}$
- maximiza a verosimilhança $\sum_{j=1}^{n} \log h_{i=y_j}(x_j)$
- treina em conjunto todos as funções g_i

Convolutional Neural Networks

Cun et. al. Gradient-based learning applied to document recognition, 1998.



Convolutional Neural Networks



- convolution: aplica filtros para detectar texturas
- strides: como avança na imagem
- rectification: usualmente aplica-se ReLU
- pooling: diminui a quantidade de pixeis

Convolutional Neural Networks