Reconhecimento de Padrões II

Tiago Buarque Assunção de Carvalho

16 de agosto de 2016

Aprendizagem de Máquina (AM)

Objetivo: prever, valores sobre elementos desconhecidos até o momento, a partir de um modelo que foi construído com um conjunto de dados similares.

Tarefas principais: Classificação, Regressão, Agrupamento, Redução de dimensionalidade.

Classificação

Prevê rótulos em um conjunto de categorias.

Ex. 1: quem é a pessoa na foto entre as pessoas cadastradas na base?

Ex. 2: este trecho da imagem representa uma face de uma pessoa?

Ex. 3: este trecho da imagem representa uma imagem de um peixe?

Ex. 4: o histórico financeiro desta pessoa representa um bom-pagador ou um mau-pagador?

Regressão

Prevê um valor numérico.

Ex. 1: qual será o preço das ações do Google na bolsa de NY em outubro de 2016?

Ex. 2: qual a quantidade de chuva em Garanhuns no inverno de 2017?

Os exemplo anteriores representam um caso especial de regressão chamado de séries temporais, neste caso os valores preditos dependem dos valores anteriores.

Ex. 3: qual o peso do peixe a partir da sua altura e comprimento?

Ex. 4: qual o preço de um imóvel a partir da sua área, bairro e índice de qualidade de acabamento?

Tarefas supervisionadas

As tarefas de classificação e regressão são ditas supervisionadas, pois os modelos (classificadores) são treinados a partir de um conjunto de dados no qual os rótulos das classes (ou valores de regressão) são conhecidos. É possível calcular erros de estimação.

Atributos (Variáveis)

Nas tarefas descritas anteriormente os valores são estimados a partir de um conjunto de outros valores. Na Estatística é comum dividir as variáveis do problema em Variáveis Dependentes (que serão estimadas) e Variáveis Independentes (que são conhecidas mesmo para os novos exemplos). No caso de classificar um histórico financeiros, a variável dependente é a saída desejada (bom ou mau pagador) e as variáveis independentes são os dados do histórico (salário, dívidas passadas etc.). As variáveis também são chamadas atributos ou características (*features*) em AM.

Conjunto de Treinamento

O conjunto de treino pode ser representado como uma matriz X, na qual cada linha $\mathbf{x}_k = (x_{k1}, \ldots, x_{kd})$, $k = 1, \ldots, n$, é um entre n exemplos, e cada coluna $\mathbf{v}_i = (x_{1i}, \ldots, x_{ni})$, $i = 1, \ldots, d$, é uma entre d variáveis. No exemplo abaixo temos $n = 5$ exemplos na base de treino e $d = 3$ dimensões.[1]

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \\ x_{51} & x_{52} & x_{53} \end{bmatrix}. \tag{1}$$

Em tarefas supervisionadas pode-se escrever o conjunto de treinamento com $C = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, em que y_k, $k = 1, \ldots, n$, é o rótulo da classe associada a $\mathbf{x}_k$. Ou seja, um conjunto no qual cada elemento é um par vetor de atributo e classe. Outra possível forma de descrever o conjunto de treino em um problema supervisionado é:

$$X' = \begin{bmatrix} x_{11} & x_{12} & x_{13} & y_1 \\ x_{21} & x_{22} & x_{23} & y_2 \\ x_{31} & x_{32} & x_{33} & y_3 \\ x_{41} & x_{42} & x_{43} & y_4 \\ x_{51} & x_{52} & x_{53} & y_5 \end{bmatrix}. \tag{2}$$

Atributos categóricos

São aqueles cujos valores pertencem a um conjunto finitos de possibilidades e que permitem apenas a operação de igualdade (=). Na classificação o classe é um atributo categórico.

Ex. 1: cor da pele segundo o IBGE = branco, amarelo, pardo, negro.

Ex. 2: bom pagador ou mau pagador (apenas dois valores possíveis).

Atributos numéricos

São aqueles cujos valores possíveis são números reais, é possível realizar além da operação de igualdade (=), as operações de comparação de ordem (<), soma (+), e multiplicação (x). Na regressão a variável da classe é numérica.

Exs.: peso, altura, salário, valore de intensidade de um pixel.

Outros tipos de atributos

Os atributos numéricos e categóricos são os mais comuns entre os modelos de AM. Mas existem outros tipos, entre eles estão os atributos ordinais e os atributos intervalares. Os atributos ordinais são aqueles que permitem operação de igualdade (=) e de comparação (<), e.g., posição de um atleta no ranking mundial. Atributos intervalares permitem operação de igualdade (=), as operações de comparação de ordem (<) e soma (+), e.g., temperatura.

Tarefas não supervisionadas

São aquelas que não tem valores alvo e não é possível definir um erro de estimação. Uma das principais tarefas não supervisionadas é o agrupamento. Outras tarefas, como redução de dimensionalidade podem ser ou não supervisionadas.

Agrupamento

Forma grupos de forma que cada grupo contém elementos que são semelhantes entre si de acordo com um critério estabelecido.

[1] Observe que o x está em negrito em $\mathbf{x}_k$, indicado que este é um vetor.

Ex. 1: taxonomia de animais (mamíferos, aves, repteis e peixes), este exemplo mostra que é difícil formar grupos perfeitos, pois alguns elementos parecem não se encaixar nestes grupos, e.g., ornitorrinco.

Ex. 2: grupos de disciplinas, formado a partir dos estudantes matriculados.

Ex. 3: grupos de atributos.

Agrupamentos Rígidos × Agrupamentos Difusos

Nos agrupamentos rígidos (*hard clusters*) cada elemento pertence a um único grupo. Nos agrupamentos difusos (*fuzzy clusters*) cada elemento pertence a todos os grupo, sendo possui um grau de pertinência alto em relação a alguns grupos e baixo em relação a outros.

Redução de dimensionalidade

O número de atributos em um problema de AM é chamado também de a Dimensão do problema. A redução de dimensionalidade consiste em reduzir a dimensão do problema sem alterar propriedades do mesmo, exemplos: sem reduzir a taxa de acerto na classificação, formando os mesmos grupos no agrupamento. Existem dois tipos principais de redução de dimensionalidade: seleção e extração de características.

Ex.: suponha um problema que recebe com entrada uma foto de uma pessoa para classificar o biotipo como "estreita" (altura muito maior que largura) ou "achatada". Se o sistema recebe como entrada uma foto de 2.0 Mega Pixel (2 milhões de pixels) em tons de cinza, e a intensidade de cada pixel é uma característica, o problema tem 2 milhões de dimensões. É possível extrair apenas duas dimensões relevantes para o problema: aquelas que tenham informação sobre a altura e a largura do indivíduo na foto.

Seleção de características

Consistem em remover alguns dos atributos mantendo apenas um conjunto selecionado. Na tarefas de classificação a seleção pode ser feita mantendo as características que maximizem a taxa de acerto na classificação.

Extração de características

Também chamado de geração de características consiste em calcular novas características a partir do conjunto inicial. A extração pode classificada em tipos: linear ou não-linear. A extração linear calcula uma nova característica com uma transformação linear dos dados (soma ponderada). Existem várias vantagens em se usar um método linear para extrair características: é inversível (é possível reconstruir os dados originais com facilidade); é possível transformar uma sequencia de operações lineares em uma única operação linear. Por outro lado a extração não-linear é mais versátil, permitindo mais operações. Um dos métodos mais conhecidos de extração de características é o PCA (*Principal Component Analysis*, análise dos componentes principais). PCA é um método linear.

PCA

Este é um método não supervisionado de extração de características para dados cujas variáveis independentes são numéricas. Quando os dados possuem atributos não-numéricos é preciso convertê-los para numéricos. O objetivo do PCA é encontrar direções ortogonais (perpendiculares) entre si com máxima variância no conjunto de dados. Entre os conceitos necessários para utilizar PCA estão: média, variância, covariância, matriz de covariância, autovalores, autovetores, produtos interno e projeção de um ponto em uma reta.

A média $\bar{v}_i$ pode ser estimada somando-se todos os valores de uma variável $\mathbf{v}_i$ no conjunto de treino e dividindo pelo número de amostras:

$$\bar{v}_i = \frac{1}{n} \sum_{k=1}^{n} x_{ki} \tag{3}$$

A variância s_i^2 é uma medida de dispersão em relação à média de uma variável $\mathbf{v}_i$:

$$s_i^2 = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \bar{v}_i)^2. \tag{4}$$

$s_i = \sqrt{s_i^2}$ é o desvio padrão. A variância e a média são estatísticas (medidas) enviesadas, isto é, podem ter seu valor significativamente alterado devido a poucos exemplos espúrios (*outiers*).

Considere as duas variáveis $\mathbf{v}_1 = \{1, 2, 3, 4, 5\}$ e $\mathbf{v}_2 = \{-3, 0, 3, 6, 9\}$. As duas variáveis têm a mesma média $\bar{v}_1 = 3$ e $\bar{v}_2 = 3$. Mas possuem variância distinta

$$s_1^2 = ((1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2)/5 = (4+1+0+1+4)/5 = 2,$$

$$s_2^2 = ((-3-3)^2 + (0-3)^2 + (3-3)^2 + (6-3)^2 + (9-3)^2)/5 = (36+9+0+9+36)/5 = 18.$$

Quando uma variável tem variância pequena, isto significa que seus valores não mudam muito de exemplo para exemplo. Pois cada valor assumido por uma variável no conjunto de treino é o valor desta para um exemplo específico. Quando uma variável tem variância alta, isto significa que os exemplos estão muito espalhados, e hipoteticamente é mais fácil de separá-los nas taredas de AM. Por esta razão o objetivo do PCA é encontrar as direções no espaço onde as variáveis estão mais espalhadas, ou seja, possuem maior variância. Exemplo, considere que a atura da exposição do olho seja uma característica para classificação de faces, para uma base de pessoas japonesas esta característica pode ter uma variância muito baixa e não ser útil para classificação.

A covariância s_{ij} mede a relação linear entre duas variáveis $\mathbf{v}_i$ e $\mathbf{v}_j$.

$$s_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \bar{v}_i)(x_{kj} - \bar{v}_j). \tag{5}$$

A covariância pode ser positiva (alta) quando ambas as variáveis comportam-se da mesma maneira, pode ser negativa (alta) quando são inversamente proporcionais, ainda pode ser próxima de zero quando não há relação linear entre as variáveis. Exemplo: $\mathbf{v}_1 = (1, 2, 3)$, $\mathbf{v}_2 = (10, 20, 30)$, $\mathbf{v}_3 = (3, 2, 1)$, $\mathbf{v}_4 = (-1, 2, 1)$; $s_{1,2} = 6,7$, $s_{1,3} = -0,67$, $s_{1,4} = 0$.

A matriz de covariância é uma matriz $S_{d\times d} = [s_{ij}]$ em que na posição (i, j) contém a covariância entre i-ésima e j-ésima variáveis, $i = 1, \ldots, d$, $j = 1, \ldots, d$ e d é o número de variáveis (dimensão) do problema.

Projetar consiste em calcular em qual posição de uma reta está um determinado ponto em um espaço d-dimensional, como pode ser visto na Figura 1. Para tanto é traçada uma linha perpendicular a reta e que passe pelo ponto em questão, o valor da projeção é o ponto no qual a linha cruza a reta.

Um vetor d-dimensional, representa um exemplo do conjunto de treino, quando este exemplo é projetado em uma reta obtêm-se um valor (um escalar). Quando projeta-se todos os n pontos do conjunto de treino, obtêm-se n valores. Estes valores são os valores de uma nova variável extraída e pode-se calcular a variância desta nova variável.

Para o primeiro componente principal encontrado com PCA, esta direção é aquela que tem a maior variância possível no conjunto de dados. O segundo componente principal é perpendicular ao primeiro e possui a maior variância possível dentro da restrição de ortogonalidade assumida. O terceiro componente principal é ortogonal aos anteriores e tem a maior variância possível satisfazendo esta restrição. E assim por diante.

Como calcular os componentes principais? Os componentes principais são os autovetores da matriz de covariância (das variáveis do conjunto de treino). Seja S uma matriz, $\mathbf{u}$ é um autovetor e λ um autovalor, se:

$$S\mathbf{u} = \lambda\mathbf{u}. \tag{6}$$

Exemplo:

$$S = \begin{bmatrix} 0.6857 & -0.0424 & 1.2743 & 0.5163 \\ -0.0424 & 0.1900 & -0.3297 & -0.1216 \\ 1.2743 & -0.3297 & 3.1163 & 1.2956 \\ 0.5163 & -0.1216 & 1.2956 & 0.5810 \end{bmatrix} \begin{bmatrix} 0.3614 \\ -0.0845 \\ 0.8567 \\ 0.3583 \end{bmatrix} = 4.2282 \begin{bmatrix} 0.3614 \\ -0.0845 \\ 0.8567 \\ 0.3583 \end{bmatrix}$$

Cada autovetor é um compomentes principal, mas qual? Qual o primeiro componente principal? Cada autovetor está associado a um autovalor, este autovalor é equivalente à variância da nova variável projetada na direção definida pelo autovetor. Assume-se que o autovetor está normalizado, ou seja, tem comprimento igual a 1. O primeiro autovetor é o que tem o maior autovalor, ou seja, a maior variância para uma nova variável projetada neste vetor. O segundo autovetor é o que tem o segundo maior autovalor e assim por diante.

Para projetar um exemplo $\mathbf{x}_k$, d-dimensional, em um vetor unitário (norma = 1), $\mathbf{u}$, basta calcular o produto interno entre estes dois vetores. Seja $\mathbf{u} = (u_1, \ldots, u_d)$ um autovetor, e $\mathbf{x}_k = (x_{k1}, \ldots, x_{kd})$ um exemplo, assumindo que $\mathbf{u}$ está normalizado, para projetar $\mathbf{x}_k$ em $\mathbf{u}$ basta calcular o produto interno entre estes dois vetores:

$$\mathbf{x}_k \cdot \mathbf{u} = \sum_{i=1}^{d} x_{ki} u_i. \tag{7}$$

Uma abordagem comum no PCA é centralizar os dados em torno da média antes da projeção, o que resultaria na seguinte equação de projeção:

$$x'_k = \sum_{i=1}^{d} (x_{ki} - \bar{v}_i) u_i. \tag{8}$$

Roteiro de projeção com PCA:

1. Calcular a matriz de covariância do conjunto de treino;
2. Calcular os autovetores e autovalores da matriz de covariância;
3. Ordenar os autovetores a partir dos autovalores de forma que $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_d$;
4. Extrair características (projetar) dos exemplos (de treino ou teste) utilizando os primeiros (quantos quiser) componentes principais utilizando a Equação 8.

Exercício

1. Baixar a base Iris <https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data>.
2. Utilizar os 10 primeiros exemplos de cada classe como a base de teste (30 exemplos de teste) e os outros 120 exemplos como conjunto de treino.
3. Utilizar o classificador 1-NN com distância Euclidiana.
4. Calcular a taxa de acerto utilizando todas as quatro dimensões.
5. Selecionar duas dimensões quaisquer para plotar o gráfico (conjunto de treino) utilizando um símbolo (ou cor) diferente para cada classe.
6. Calcular a taxa de acerto utilizando apenas as duas dimensões selecionadas.
7. Treinar o PCA no conjunto de treino.

8. Utilizar as médias e os dois componentes principais calculados no conjunto de treino, projetar os conjuntos de treino e teste.

9. Gerar o gráfico para o conjunto de treino com as duas dimensões extraídas com PCA, diferenciando cada classe.

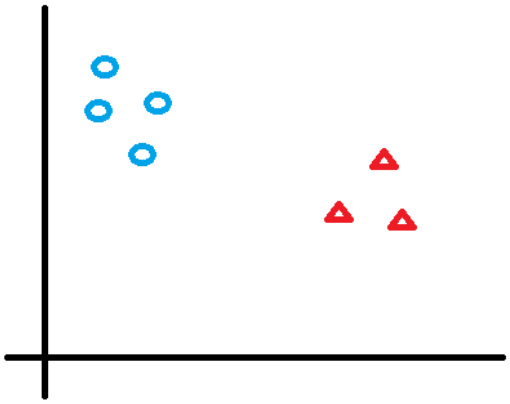
10. Gerar o gráfico para o conjunto de teste com as duas dimensões extraídas com PCA, diferenciando cada classe.

11. Calcular a taxa de acerto utilizando as duas características extraídas com PCA.

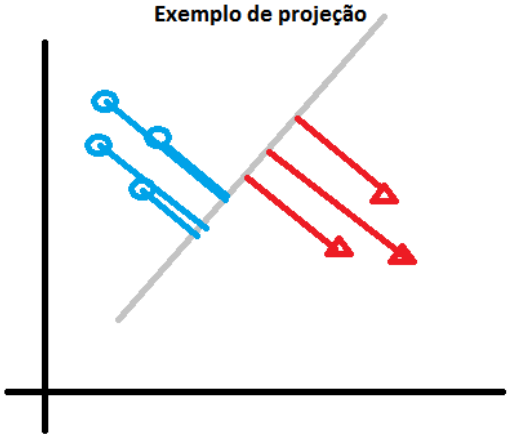
Experimento em Aprendizagem de Máquina

Referências bibliográficas

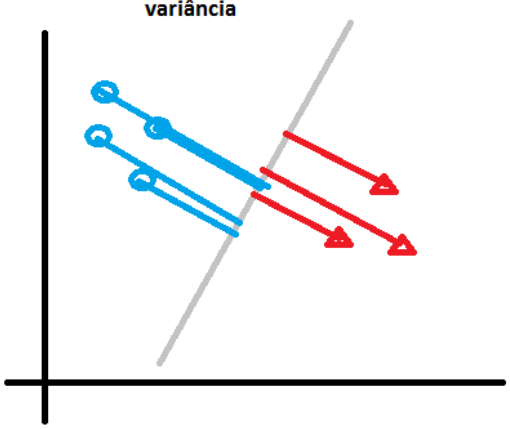
Exemplo de dados



Exemplo de projeção



Projeção de baixa variância



PCA
Direção de maior variância
= Componente principal