Advanced Institute for Artificial Intelligence

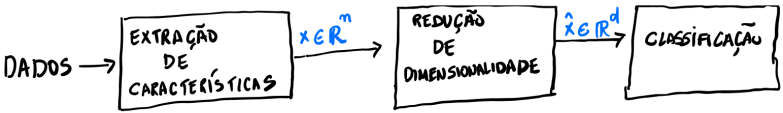
Análise de Componentes Principais

Advanced Institute for Artificial Intelligence – AI2

https://advancedinstitute.ai

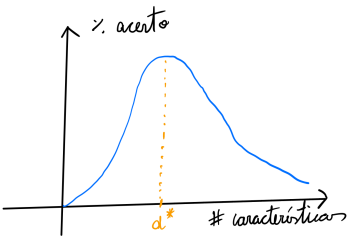
Introdução

Técnicas de redução de dimensionalidade/transformação do espaço de características visam obter versões mais **compactas/representativas** de nossos dados. Dado um conjunto de dados $\mathcal{X} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_z, y_z)\}$ em que $\boldsymbol{x}_i \in \mathbb{R}^n$ e $y_i \in \mathbb{N}$, a ideia consiste em obter um novo conjunto $\hat{\mathcal{X}} = \{(\hat{\boldsymbol{x}}_1, y_1), (\hat{\boldsymbol{x}}_2, y_2), \ldots, (\hat{\boldsymbol{x}}_z, y_z)\}$, em que $\hat{\boldsymbol{x}}_i \in \mathbb{R}^d$ e $d < n$.

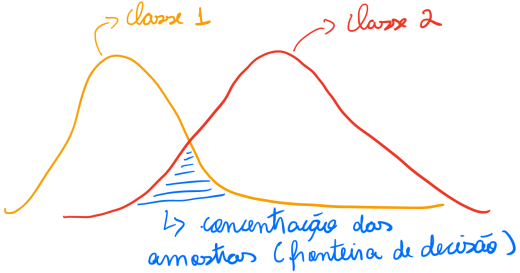


Usualmente, em problemas de classificação, temos a impressão de, quanto mais características temos, melhor será a taxa de acerto de nossa técnica. No entanto, em problemas reais em que temos um número **limitado de amostras**, observa-se um fenômeno conhecido por **maldição da dimensionalidade.** Existe, então, uma série de efeitos negativos ocasionados pelo aumento indiscriminado de características:

- Fenômeno de Hughes: para um número finito de amostras, existe uma dimensionalidade d^* que, após este valor, o desempenho da taxa de classificação diminui.



- Número de amostras como função das características: em classificadores não paramétricos, o número de amostras deve ser uma função exponencial do número de características.
- Gaussianas multivariadas: em distribuições Gaussianas multivariadas com alta dimensão, a densidade das amostras tende a se concentrar na cauda da distribuição, ou seja, longe da média amostral, dificultando a classificação.



Autovalores e Autovetores: Uma Breve Introdução

Seja V um espaço vetorial com n dimensões que contempla um produto interno e dois vetores $\boldsymbol{u}, \boldsymbol{v} \in V$. Supondo que V seja um espaço Euclidiano, temos que seus vetores possuem uma **direção** e uma **magnitude**. O **produto interno** entre seus vetores pode ser calculado da seguinte forma:

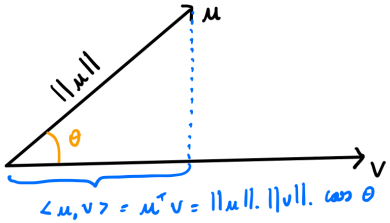
$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^T \boldsymbol{v} = \sum_{i=1}^{n} u_i v_i. \tag{1}$$

Uma outra possibilidade para calcular o produto interno entre dois vetores é dada como segue:

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \|\boldsymbol{u}\|.\|\boldsymbol{v}\|.\cos(\theta), \tag{2}$$

em que θ corresponde ao ângulo entre os dois vetores e $\|\cdot\|$ representa a magnitude do vetor.

Esta formulação modela o produto interno entre os vetores $\boldsymbol{u}$ e $\boldsymbol{v}$ como sendo a **projeção** de $\boldsymbol{u}$ em $\boldsymbol{v}$.



Temos, também, que:

$$\langle \boldsymbol{v}, \boldsymbol{v} \rangle = \boldsymbol{v}^T \boldsymbol{v} = \sum_{i=1}^{n} v_i^2 = \|v\|^2. \tag{3}$$

Existem diferentes maneiras de calcular a norma (magnitude) de um vetor. A norma Euclidiana $\|\boldsymbol{v}\|_2$ de um vetor, também chamada de norma L_2, é calculada da seguinte forma:

$$\|\boldsymbol{v}\|_2 = \sqrt{\sum_{i=1}^{n} v_i^2}. \tag{4}$$

Usualmente, denotamos a norma Euclidiana por $\|\cdot\|$.

A ideia é que neste espaço vetorial eu consiga definir **operadores lineares**, ou seja, funções (matrizes) $\boldsymbol{P}$ que mapeiam vetores de entrada $\boldsymbol{v}$ em vetores de saída μu, ou seja:

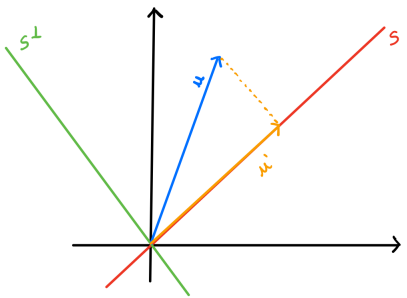
$$\boldsymbol{u} = \boldsymbol{P}\boldsymbol{v}. \tag{5}$$

No estudo de autovalores e autovetores, estamos interessados em operadores lineares que possuem a seguinte característica: dado um operador linear $\boldsymbol{P}$ para quais vetores $\boldsymbol{v}$ a saída da Equação 3, ou seja, $\boldsymbol{u}$, aponta para a mesma direção da entrada, sendo apenas esticados ou encolhidos? Matematicamente falando, temos que:

$$\boldsymbol{u} = \boldsymbol{P}\boldsymbol{v} = \lambda \boldsymbol{v}. \tag{6}$$

Todos os vetores $\boldsymbol{v}$ que satisfazem a equação acima são chamados de **autovetores** de $\boldsymbol{P}$, e todos os escalares λ que satisfazem a formulação acima são chamados de **autovalores** de $\boldsymbol{P}$. Na prática, λ é um fator de escala de redução ou incremento do vetor $\boldsymbol{v}$. Assim sendo, dizemos que $\boldsymbol{v} \in \mathbb{R}^n$ é um autovetor de $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ com autovalor λ se $\boldsymbol{v} \neq 0$ e $\boldsymbol{P}\boldsymbol{v} = \lambda \boldsymbol{v}$.

Para fins de explicação, tomemos o seguinte exemplo: seja $\boldsymbol{P} = \boldsymbol{I} \in \mathbb{R}^{n \times n}$ o operador identidade. Então, $\boldsymbol{I}\boldsymbol{v} = 1\boldsymbol{v}$, $\forall \boldsymbol{v} \in \mathbb{R}^n$. Desta forma, $\boldsymbol{v}$ é um autovetor de $\boldsymbol{I}$ com autovalor $\lambda = 1$. Vejamos um outro exemplo, o operador de projeção em $\mathbb{R}^2$.



Podemos definir o operador $\boldsymbol{P}_S$, o qual projeta o vetor $\boldsymbol{u}$ no subespaço (reta) S:

$$\boldsymbol{u}' = \boldsymbol{P}_S \boldsymbol{u}. \tag{7}$$

Supondo um vetor $\boldsymbol{w} \in S$, temos que $\boldsymbol{w} = \boldsymbol{P}_S \boldsymbol{w}$, dado que $\boldsymbol{w}$ já faz parte do subespaço S. Neste caso, todo vetor que pertence à reta S é um autovetor de $\boldsymbol{P}_S$ com autovalor $\lambda = 1$.

Seja, agora, $S^\perp$ o subespaço ortogonal a $\boldsymbol{S}$, e $\boldsymbol{x} \in S^\perp$, ou seja, quando eu aplicar o operador $\boldsymbol{P}_S$ em algum elemento de $\boldsymbol{S}^\perp$, resulta no valor 0 (origem). Então, temos que $\boldsymbol{P}_S \boldsymbol{x} = 0\boldsymbol{x}$, ou seja, todo vetor que pertence a $S^\perp$ é um autovetor de $\boldsymbol{P}_S$ com autovalor $\lambda = 0$.

Podemos extrapolar esse arcabouço de autovetores e autovalores para matrizes. Por exemplo, como podemos calcular os autovalores e autovetores de uma matrix $\boldsymbol{A}$? Da Equação 6, temos que:

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}.$$

Rearranjando os temos, temos:

$$\boldsymbol{A}\boldsymbol{v} - \lambda \boldsymbol{I}\boldsymbol{v} = 0 \implies (\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{v} = 0. \tag{8}$$

Existe um teorema da álgebra linear que nos diz o seguinte: $\boldsymbol{B}\boldsymbol{v} = 0$ admite solução não nula se, e somente se, $det(\boldsymbol{B}) = 0$. Assim, $det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$ em nosso caso.

Ex: seja a matriz:

$$\boldsymbol{A} = \begin{bmatrix} 3/2 & -1/2 \\ -1/2 & 3/2 \end{bmatrix}.$$

Temos que:

$$\begin{aligned} \boldsymbol{A} - \lambda \boldsymbol{I} &= \begin{bmatrix} 3/2 & -1/2 \\ -1/2 & 3/2 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 3/2 - \lambda & -1/2 \\ -1/2 & 3/2 - \lambda \end{bmatrix}. \end{aligned}$$

Sabemos que $det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$, ou seja:

$$\left(\frac{3}{2} - \lambda\right)^2 - \left(-\frac{1}{2} * -\frac{1}{2}\right) = \left(\frac{3}{2} - \lambda\right)^2 - \frac{1}{4} = 0,$$

o que implica em:

$$\frac{9}{4} - 2\frac{3}{2}\lambda + \lambda^2 - \frac{1}{4} = 0 \implies \lambda^2 - 3\lambda + 2 = 0.$$

Assim sendo, temos uma equação do segundo grau cujas soluções são $\lambda_1 = 1$ e $\lambda_2 = 2$.

Para obtermos o autovetor associado ao autovalor $\lambda_1 = 1$, temos que:

$$\begin{aligned} \boldsymbol{A} - \lambda \boldsymbol{I} &= \begin{bmatrix} 3/2 & -1/2 \\ -1/2 & 3/2 \end{bmatrix} - 1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 3/2 - 1 & -1/2 \\ -1/2 & 3/2 - 1 \end{bmatrix} \\ &= \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}. \end{aligned} \tag{9}$$

Continuando, substituindo o resultado da Equação 7 na Equação 6, temos:

$$(\boldsymbol{A} - \lambda \boldsymbol{I})\boldsymbol{v} = 0$$

$$\begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix} \boldsymbol{v} = 0$$

$$\begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \implies \begin{cases} \frac{1}{2}v_1 - \frac{1}{2}v_2 = 0 \\ -\frac{1}{2}v_1 + \frac{1}{2}v_2 = 0 \end{cases} \implies v_1 = v_2 \implies \boldsymbol{v}^{(1)}$$

Desta forma, $\boldsymbol{v}^{(1)}$ é o autovetor associado ao autovalor λ_1.

Note que temos infinitos autovetores associadores ao autovalor λ_1, basta apenas que as componentes tenham o mesmo valor. Ex: $\boldsymbol{v}^{(1)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ ou $\boldsymbol{v}^{(1)} = \begin{bmatrix} 7 \\ 7 \end{bmatrix}$. Qual a diferença entre eles?

Todos apontam para a mesma direção, mas possuem magnitudes diferentes.

Usualmente, utilizamos o **autovetor canônico**, ou seja, aquele que tem **norma unitária**. Basta tomarmos um autovetor qualquer e dividirmos cada componente pela sua norma, ou seja:

$$\boldsymbol{v}^{(1)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \implies \boldsymbol{v}^{(1)} = \begin{bmatrix} \frac{1}{\|\boldsymbol{v}^{(1)}\|} \\ \frac{1}{\|\boldsymbol{v}^{(1)}\|} \end{bmatrix} \implies \boldsymbol{v}^{(1)} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}.$$

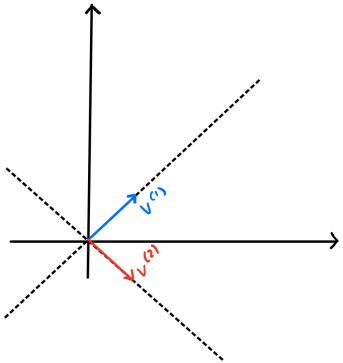
Agora, repetimos o processo para o autovalor $\lambda_2 = 2$:

$$\begin{bmatrix} -1/2 & -1/2 \\ -1/2 & -1/2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \implies \begin{cases} -\frac{1}{2}v_1 - \frac{1}{2}v_2 = 0 \\ -\frac{1}{2}v_1 - \frac{1}{2}v_2 = 0 \end{cases} \implies v_1 = -v_2 \implies \boldsymbol{v}^{(2)}.$$

Novamente, utilizamos o autovetor canônico:

$$\boldsymbol{v}^{(2)} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \implies \boldsymbol{v}^{(2)} = \begin{bmatrix} \frac{1}{\left\|\boldsymbol{v}^{(1)}\right\|} \\ -\frac{1}{\left\|\boldsymbol{v}^{(1)}\right\|} \end{bmatrix} \implies \boldsymbol{v}^{(2)} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

Desta forma, temos que $\boldsymbol{v}^{(1)}$ e $\boldsymbol{v}^{(2)}$ são os nossos autovetores do operador linear $\boldsymbol{P}_S$. Geometricamente falando, temos o seguinte:



Temos que os autovetores definem uma nova base no sistema de coordenadas. Existe um teorema que nos diz o seguinte: "Se $\boldsymbol{P}$ é um operador linear que possui n autovalores distintos, então os autovetores de $\boldsymbol{P}$ definem uma nova base em $\mathbb{R}^n$."

Podemos organizar os autovetores em uma matriz $\boldsymbol{Q}$ da seguinte forma:

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{v}^{(1)} & \boldsymbol{v}^{(2)} & \dots & \boldsymbol{v}^{(n)}. \end{bmatrix}$$

Ademais, temos que uma matriz é dita ser **positiva semidefinida** se todos os seus autovalores foram maiores ou iguais a 0. Temos, ainda, uma outra definição: "Seja $\boldsymbol{P}$ um operador linear. Caso a matriz $\boldsymbol{Q}$ dos seus autovetores defina uma nova base em $\mathbb{R}^n$, então $\boldsymbol{PQ} = \boldsymbol{Q\Lambda}$, em que $\boldsymbol{\Lambda}$ representa a matriz diagonal dos autovalores".

Um outro teorema fundamental (decomposição expectral) nos diz que: "Seja $\boldsymbol{P}$ uma matriz quadrada $n \times n$, $\boldsymbol{Q}$ a sua matriz de autovetores (nas colunas) e $\boldsymbol{\Lambda}$ a sua matriz diagonal de autovalores. Então, temos que a matriz $\boldsymbol{P}$ pode ser decomposta da seguinte forma:"

$$\boldsymbol{P} = \boldsymbol{Q\Lambda Q}^{-1}. \tag{10}$$

Caso $\boldsymbol{P}$ seja ortogonal (ex: matrizes de rotação do espaço), então $\boldsymbol{Q}^{-1} = \boldsymbol{Q}^T$. Desta forma, temos que $\boldsymbol{P} = \boldsymbol{Q\Lambda Q}^T$.

Análise de Componentes Principais pela Maximização da Variância

A técnica PCA nos permite tratar dados de alta dimensionalidade identificando a dependência entre as variáveis para representá-los de uma forma mais compacta e minimizando a perda de informação relevante. É uma das técnicas mais utilizadas no contexto de **redução de características**. Possui outros nomes, tais como: (i) Transformação de Karhunen-Loeve, (ii) Transformação de Hotelling e (iii) Decomposição em Valores Singulares.

Características principais:

- Transformação linear: assume a hipótese que os dados encontram-se em um subespaço Euclidiano de $\mathbb{R}^n$.
- Método não supervisionado.
- Decorrelaciona os dados de entrada eliminando redundâncias (a matriz de covariância dos dados transformados é diagonal).

$$\boldsymbol{x} \in \mathbb{R}^n \overset{\text{PCA}}{\Longrightarrow} \hat{\boldsymbol{x}} \in \mathbb{R}^d$$

Basicamente, queremos achar um operador que projete os dados de entrada em um espaço de saída com menor dimensão.

Seja $Z = [T^T, S^T]$ uma base ortonormal em $\mathbb{R}^n$, em que:

- $T^T = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_d]$ corresponde ao vetor de componentes que iremos **reter** no processo de redução e
- $S^T = [\boldsymbol{w}_{k+1}, \boldsymbol{w}_{k+2}, \ldots, \boldsymbol{w}_n]$ corresponde ao vetor de componentes que iremos **descartar**.

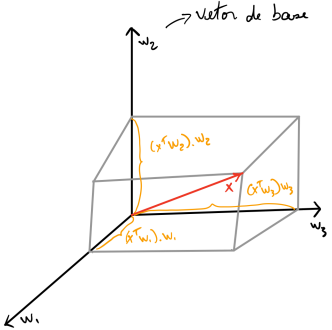
Desta forma, T representa o novo sistema de eixos coordenados encontrados pelo PCA, enquanto que S representa o subespaço eliminado durante o processo de redução.

Problema: dado um espaço de entrada, queremos encontrar as d direções $\boldsymbol{w}_i$ que, ao projetarmos os dados, maximizam a variância (espalhamento) retida na nova representação.

Temos que nosso vetor $\boldsymbol{x} \in \mathbb{R}^n$ pode ser expandido em nossa base ortonormal da seguinte forma:

$$\boldsymbol{x} = \sum_{j=1}^{n} (\boldsymbol{x}^T \boldsymbol{w}_j) \boldsymbol{w}_j = \sum_{j=1}^{n} c_j \boldsymbol{w}_j, \tag{11}$$

em que $\boldsymbol{c} \in \mathbb{R}^n$ corresponde ao vetor de coeficientes de expansão.



Note que o termo $\boldsymbol{x}^T \boldsymbol{w}_j$ corresponde à projeção de $\boldsymbol{x}$ no vetor de base $\boldsymbol{w}_j$.

Temos que o vetor reduzido para o novo espaço pode ser calculado da seguinte forma:

$$\hat{\boldsymbol{x}} = T\boldsymbol{x} \implies \hat{\boldsymbol{x}}^T = \boldsymbol{x}^T T^T = \sum_{j=1}^{n} c_j \boldsymbol{w}_j^T T^T. \tag{12}$$

Note que a propriedade de ortonormalidade (bases ortonormais) nos diz que:

$$\boldsymbol{w}_i^T \boldsymbol{w}_j = \begin{cases} 1 & \text{se } i = j \\ 0 & \text{caso contrário.} \end{cases} \tag{13}$$

Assim, podemos reescrever a Equação 10 conforme segue:

$$\hat{\boldsymbol{x}} = \sum_{j=1}^{n} c_j \boldsymbol{w}_j^T T^T = \sum_{j=1}^{n} c_j \boldsymbol{w}_j^T [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_d] = [c_1, c_2, \ldots, c_d]. \tag{14}$$

Assim, queremos encontrar uma transformação T (conjunto de componentes) que **maximize** a variância retida nos dados, ou seja, queremos maximizar a seguinte função:

$$L(T) = E\left[\|\hat{\boldsymbol{x}}\|^2\right] = \left[\hat{\boldsymbol{x}}^T\hat{\boldsymbol{x}}\right] = \sum_{j=1}^{d} E[c_j^2], \tag{15}$$

lembrando que $\left\|\hat{\boldsymbol{x}}^2\right\|$ corresponde à norma ao quadrado de $\hat{\boldsymbol{x}}$, ou seja, queremos obter novas componentes cujas normas (distâncias a partir da origem) sejam maximizadas, ou seja, estejam mais "espalhadas". Note que $\left[\hat{\boldsymbol{x}}^T\hat{\boldsymbol{x}}\right] = E\left[\hat{x}_1^2\right] + E\left[\hat{x}_2^2\right] + \ldots + E\left[\hat{x}_d^2\right]$, ou seja, a soma das variâncias em cada eixo coordenado da nova representação.

Da Equação 9, temos que $c_j = \boldsymbol{x}^T \boldsymbol{w}_j$. Assim, substituindo-se c_j na Equação 15, temos que:

$$\begin{aligned} L(T) &= \sum_{i=1}^{d} E[c_i^2] = \sum_{i=1}^{d} E[(\boldsymbol{x}^T \boldsymbol{w}_i)^2] \\ &= \sum_{i=1}^{d} E[\boldsymbol{w}_i^T \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{w}_i] = \sum_{i=1}^{d} \boldsymbol{w}_i^T E[\boldsymbol{x}\boldsymbol{x}^T] \boldsymbol{w}_i \\ &= \sum_{i=1}^{d} \boldsymbol{w}_i^T \Sigma_{\boldsymbol{x}} \boldsymbol{w}_i, \end{aligned} \tag{16}$$

em que $\Sigma_{\boldsymbol{x}}$ corresponde à matrix de covariância dos dados observados, ou seja, do vetor de entrada $\boldsymbol{x}$. Ademais, temos que a Equação 16 está sujeita à restrição $\|\boldsymbol{w}_i\|^2 = 1$.

Desta forma, o problema de otimização consiste em:

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} \sum_{i=1}^{d} \boldsymbol{w}_i^T \Sigma_{\boldsymbol{x}} \boldsymbol{w}_i, \tag{17}$$

sujeito à $\|\boldsymbol{w}_i\|^2 = 1$, $\forall i = 1, 2 \ldots, d$. Desta forma, nosso problema consiste em encontrar as d direções ortogonais $\boldsymbol{w}_i$ que maximizam o somatório acima. Note que o somatório contém a matriz de covariância dos dados, ou seja, queremos maximizar a sua variância nessas direções ortogonais.

Temos, então, um problema de otimização com restrições de desigualdade, ou seja, precisamos fazer uso da técnica de Multiplicadores de Lagrange.

Escrevendo a Equação 16 de acordo com os Multiplicadores de Lagrange, temos que:

$$L(T, \boldsymbol{\alpha}) = \sum_{i=1}^{d} \boldsymbol{w}_i^T \Sigma_{\boldsymbol{x}} \boldsymbol{w}_i - \sum_{i=1}^{d} \alpha_i (\underbrace{\boldsymbol{w}_i^T \boldsymbol{w}_i}_{\|\boldsymbol{w}\|^2=1} - 1). \tag{18}$$

Como queremos maximizar a equação acima visando cada $\boldsymbol{w}_i$, basta, então, calcularmos a derivada de $L(T, \boldsymbol{\alpha})$ em relação à $\boldsymbol{w}_i$ e igualarmos à 0, ou seja:

$$\frac{\partial L(T, \boldsymbol{\alpha})}{\partial \boldsymbol{w}_i} = \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{w}_i - \alpha_i \boldsymbol{w}_i = 0 \implies \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{w}_i = \alpha_i \boldsymbol{w}_i. \tag{19}$$

O resultado da equação acima nos leva à formulação do problema de encontrar os autovalores e autovetores. **Assim, as direções $\boldsymbol{w}_i$ são os autovetores da matriz de covariância dos dados de entrada.**

Podemos reescrever a Equação 17 da seguinte forma:

$$\begin{aligned} \boldsymbol{w}^* &= \arg\max_{\boldsymbol{w}} \sum_{i=1}^{d} \boldsymbol{w}_i^T \Sigma_{\boldsymbol{x}} \boldsymbol{w}_i = \arg\max_{\boldsymbol{w}} \sum_{i=1}^{d} \boldsymbol{w}_i^T \alpha_i \boldsymbol{w}_i \\ &= \arg\max_{\boldsymbol{w}} \sum_{i=1}^{d} \alpha_i \cancelto{1}{\|\boldsymbol{w}_i\|^2} = \arg\max_{\boldsymbol{w}} \sum_{i=1}^{d} \alpha_i. \end{aligned} \tag{20}$$

A equação acima nos diz que devemos maximizar o somatório dos d autovalores. Isto define o nosso problema, ou seja, encontrar as d direções que maximizam a variância dos dados!

Agora, veremos por que PCA decorrelaciona os dados, ou seja, a matriz de covariância do vetor transformado $\hat{\boldsymbol{x}}$ é uma matriz diagonal. Suponha a seguinte transformação **linear** abaixo:

$$\hat{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x}. \tag{21}$$

Existe um teorema que diz a matriz de covariância de $\hat{\boldsymbol{x}}$ pode ser obtida da seguinte forma:

$$\Sigma_{\hat{\boldsymbol{x}}} = \boldsymbol{A}^T \Sigma_{\boldsymbol{x}} \boldsymbol{A}. \tag{22}$$

Ademais, temos que toda matriz simétrica e positiva semidefinida possui uma decomposição $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$, em que $\boldsymbol{Q}$ é a matriz coluna dos autovetores e $\boldsymbol{\Lambda}$ é a matriz diagonal dos seus autovalores (**decomposição espectral**, como vimos anteriormente). **Sabemos que as matrizes de covariância são positivas e semidefinidas!**

Desta forma, podemos reescrever a Equação 22 da seguinte forma:

$$\Sigma_{\hat{\boldsymbol{x}}} = \boldsymbol{A}^T \Sigma_{\boldsymbol{x}} \boldsymbol{A} = \boldsymbol{A}^T \underbrace{\boldsymbol{Q} \boldsymbol{\Lambda} \boldsymbol{Q}^T}_{\Sigma_{\boldsymbol{x}}} \boldsymbol{A}. \tag{23}$$

Note que a matriz $\boldsymbol{A}$ na Equação 21, caso consideremos essa transformação linear sendo dada pelo PCA, acaba sendo a própria matriz de autovetores, ou seja, $\boldsymbol{A} = \boldsymbol{Q}$. Assim, podemos reescrever a Equação 23 como segue:

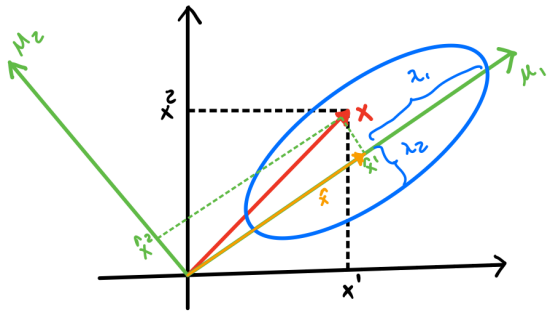
$$\Sigma_{\hat{\boldsymbol{x}}} = \boldsymbol{A}^T \boldsymbol{Q} \boldsymbol{\Lambda} \boldsymbol{Q}^T \boldsymbol{A} = \boldsymbol{Q}^T \boldsymbol{Q} \boldsymbol{\Lambda} \boldsymbol{Q}^T \boldsymbol{Q}. \tag{24}$$

Como a nossa base é **ortonormal**, temos que $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$. Assim, chegamos à uma formulação final para $\Sigma_{\hat{\boldsymbol{x}}}$:

$$\Sigma_{\hat{\boldsymbol{x}}} = \boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & \lambda_d \end{bmatrix}, \tag{25}$$

que é uma **matriz diagonal**, ou seja os dados estão decorrelacionados! Assim, após a transformação PCA, dizemos que não há correlação entre os atributos selecionados.

Vejamos, agora, uma interpretação geométrica do PCA.



Neste caso, temos que os eixos coordenados na cor preta representam o espaço original, e os eixos verdes representam o espaço transformado via PCA. Note que os novos eixos são **ortogonais** entre si. O vetor $\boldsymbol{x}$ acaba sendo projetado no eixo u_2, pois a variância nele é maior do que em u_2.

Um outro ponto interessante é que os autovalores λ_1 e λ_2 são, na verdade, as **variâncias dos novos eixos**, ou seja, eles codificam o quanto alongada está a elipse.

Análise de Componentes Principais pela Minimização do Erro Quadrático Médio

Um segundo critério para formularmos o problema de otimização do PCA é pelo erro quadrático médio. Desta forma, o PCA é ótimo em dois sentidos, isto é, ele **maximiza** o espalhamento dos dados e também **minimiza** o erro quadrático médio do vetor original para o vetor reduzido. Assim, podemos reformar a função a ser otimizada para:

$$L(T) = E\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2\right] = E\left[\left\|\boldsymbol{x} - \sum_{j=1}^{d}(\boldsymbol{w}_j^T\boldsymbol{x})\boldsymbol{w}_j\right\|^2\right]. \tag{26}$$

Aplicando a definição de norma ao quadrado, temos que:

$$L(T) = E\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2\right] = E\left[\left(\boldsymbol{x} - \sum_{j=1}^{d}(\boldsymbol{w}_j^T\boldsymbol{x})\boldsymbol{w}_j\right)\left(\boldsymbol{x} - \sum_{j=1}^{d}(\boldsymbol{w}_j^T\boldsymbol{x})\boldsymbol{w}_j\right)\right]. \quad (27)$$

Após aplicarmos operações distributivas e rearranjando os temos, temos que:

$$L(T) = E\left[\|\boldsymbol{x}\|^2\right] - \sum_{j=1}^{d} E\left[(\boldsymbol{w}_j^T\boldsymbol{x})^2\right]. \quad (28)$$

Note que o primeiro termo não depende de $\boldsymbol{w}_j$ (que é o queremos encontrar), sendo uma constante. Desta forma, para minimizar o erro quadrático médio, precisamos **maximizar** o segundo termo, pois ele é negativo. Este termo nada mais é do que a variância retida em cada novo eixo coordenado da base T.

Assim, podemos reescrever a Equação 28 como segue:

$$
\begin{aligned}
L(T) &= E\left[\|\boldsymbol{x}\|^2\right] - \sum_{j=1}^{d} E\left[(\boldsymbol{w}_j^T \boldsymbol{x})^2\right] \\
&= E\left[\|\boldsymbol{x}\|^2\right] - \sum_{j=1}^{d} E\left[(\boldsymbol{w}_j^T \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{w}_j)\right] \\
&= E\left[\|\boldsymbol{x}\|^2\right] - \sum_{j=1}^{d} \boldsymbol{w}_j^T E\left[\boldsymbol{x}\boldsymbol{x}^T\right] \boldsymbol{w}_j \\
&= E\left[\|\boldsymbol{x}\|^2\right] - \sum_{j=1}^{d} \boldsymbol{w}_j^T \Sigma_{\boldsymbol{x}} \boldsymbol{w}_j
\end{aligned}
\tag{29}
$$

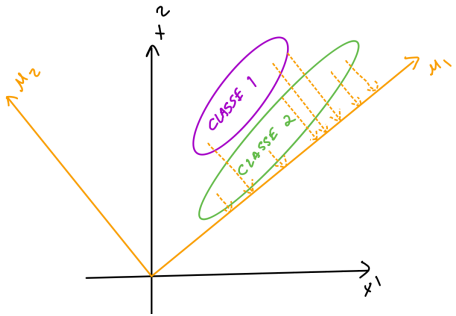
A equação anterior nos remete à mesma forma quadrática da formulação do PCA com a maximização das variâncias. **Assim sendo, para minimizarmos o erro quadrático médio, basta maximizarmos o espalhamento (variâncias) novamente!** Como fizemos anteriormente, basta derivarmos a função de custo dada pela Equação 29 com relação à $\boldsymbol{w}_j$, $\forall j = 1, 2, \ldots, d$ e igualarmos à 0. Chegaremos no mesmo caso anterior, ou seja, a solução é dada na forma de uma equação de autovalores e autovetores, isto é, $\Sigma_{\boldsymbol{x}}\boldsymbol{w}_j = \lambda_j \boldsymbol{w}_j$. Isto significa que, voltando à Equação 29, temos a seguinte equivalência:

$$
\begin{aligned}
L(T) &= E\left[\|\boldsymbol{x}\|^2\right] - \sum_{j=1}^{d} \boldsymbol{w}_j^T \Sigma_{\boldsymbol{x}} \boldsymbol{w}_j \\
&= E\left[\|\boldsymbol{x}\|^2\right] - \sum_{j=1}^{d} \boldsymbol{w}_j^T \lambda_j \boldsymbol{w}_j = E\left[\|\boldsymbol{x}\|^2\right] - \sum_{j=1}^{d} \lambda_j. \qquad (30)
\end{aligned}
$$

O último termo é simplicado por conta das propriedade de ortonormalidade.

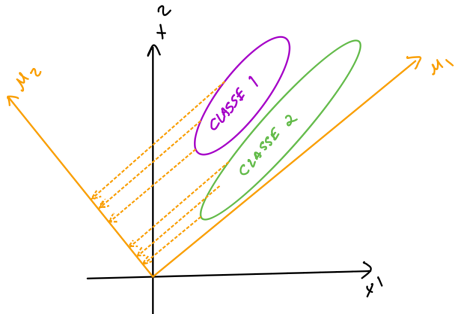
Algumas limitações do PCA:

- PCA foi projeto para **compactação** de dados, e não classificação. Caso a gente queira compactar os dados, PCA é a primeira escolha pois ele é ótimo neste sentido. No entanto, ele não é supervisionado.



Podemos observar que, nesta situação, PCA vai projetar as amostras na direção de maior variância, ou seja, u_1. No entanto, as amostras das classes diferentes ficarão todas sobrepostas.

No entanto, caso projetássemos as amostras na direção u_2 (melhor espalhamento), elas estariam melhor separadas. Esta é, então, uma limitação da técnica PCA.



Vejamos, então, o algoritmo do PCA.

1. Carregar a base de dados $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$, tal que $\boldsymbol{x}_i \in \mathbb{R}^n$.
2. Calcule o vetor média global $\boldsymbol{\mu}_{\boldsymbol{x}} \in \mathbb{R}^n$ e a matriz de covariância $\Sigma_{\boldsymbol{x}} \in \mathbb{R}^{n \times n}$.
3. Calcule os autovetores e autovalores de $\Sigma_{\boldsymbol{x}}$.
4. Selecione os d autovetores associados aos d maiores autovalores da matriz de covariância, denotados por $[\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_d]$.
5. Defina a matriz de transformação $\boldsymbol{W}_{PCA} \in \mathbb{R}^{n \times d}$.
6. Projetar dados na nova base $\hat{\boldsymbol{x}}_i = \boldsymbol{W}_{PCA}^T \boldsymbol{x}_i$, $\forall i = 1, 2, \ldots, m$.