Reconhecimento de Padrões
Introdução

Prof. George von Borries
Departamento de Estatística
Universidade de Brasília

1 - 2024

Definições e Conceitos

- **Estatística:** (*Statistics*) é uma área da matemática que concentra na construção de modelos. O objetivo é a realização de inferências, sobre a relação entre variáveis, baseadas em parâmetros estimados a partir de amostras de uma população especificada. Geralmente é aplicada a dados de baixa dimensão.
- **Reconhecimento de Padrões:** (*Pattern Recognition*) estudo de técnicas para classificação de objetos (observações) em certo número de categorias ou classes. Os objetos geralmente são imagens, sinais, doenças, etc, constituídos de medidas que precisam ser classificadas em padrões (*patterns*).

Definições e Conceitos

- **Aprendizado de Máquinas:** (*Machine Learning*) é uma área da computação ligada a inteligência artificial (*AI*) que concentra na criação de sistemas que aprendem[1] a partir dos dados. O objetivo é a otimização de predições, geralmente sem suposições sobre os dados que são de alta dimensão. A estatística, neste caso, é a base para a construção de algoritmos eficientes e não existe preocupação com interpretação dos modelos, mas com a acurácia da predição.
- **Aprendizado Estatístico:** (*Statistical Learning*) é uma área que une a estatística e computação no reconhecimento de padrões e aprendizado de máquinas. Poderíamos dizer que faz parte da ciências de dados (i.e., algo mais amplo). O foco está no conjunto de ferramentas para modelagem e compreensão de dados complexos. O objetivo é a construção de modelos estatísticos para predição e compreensão dos dados. Algum conhecimento é necessário sobre sobre a população a que os dados pertencem. Geralmente é aplicada em dados superdimensionados.

[1] Um programa computacional aprende por experiência E em relação a alguma classe de tarefas T e medida de performance P, se sua performance nas tarefas T, medidas por P, aumentam com a experiência E.
(Mitchell, T. (1997) Machine Learning. McGraw Hill)

Observação

Na realidade, estes conceitos se confundem e muitas vezes são usados para identificar as mesmas técnicas, mas em diferentes áreas de estudo. Muitas das técnicas do estudo de discriminantes, como classificação e agrupamento de dados, recebem diferentes nomes em diferentes áreas.

- Análise multivariada em Estatística.
- Reconhecimento de Padrões na Engenharia.
- Aprendizado de máquinas em Ciência da Computação.
- Aprendizado estatístico na Ciência de Dados.

Neste curso iremos tratar de tópicos em análise de discriminantes, classificação e agrupamentos, sem especificar a área de intersecção.

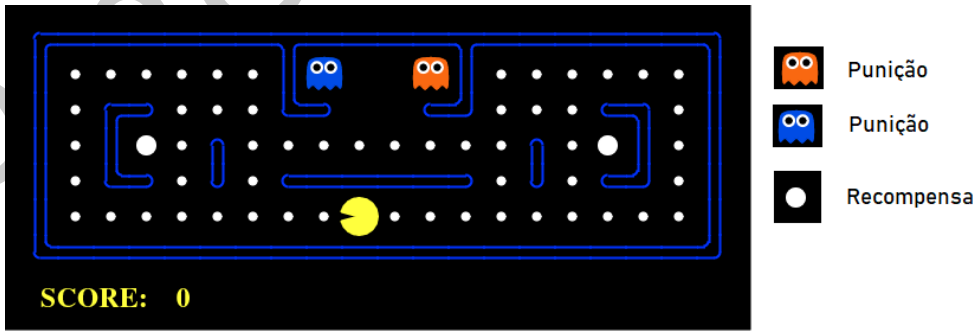
A abordagem dará enfase a métodos estatísticos.

Tipos de Aprendizado

- **Aprendizado supervisionado:** (ou preditivo) tem como objetivo aprender a relação (*mapping*) entre covariáveis $\mathbf{x}$ e respostas y, dado um conjunto de treinamento $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$. Neste caso, y_i serve como uma espécie de "supervisor" para verificação da qualidade do apredizado. Em problemas de classificação, y_i é uma variável categórica ou nominal de algum conjunto finito, i.e., $y_i \in \{1, \ldots, C\}$. A resposta pode ser também ordinal, envolvendo um nível a mais de restrição quando desejamos classificar y_i, e neste caso podemos utilizar modelos de regressão ordinal.
- **Aprendizado não supervisionado:** (ou descritivo) possui conjunto de dados de treinamento $\mathcal{T} = \{\mathbf{x}_i\}_{i=1}^{N}$ em que o objetivo é descobrir padrões no conjunto de dados. Note que neste caso não existe um "supervisor" e não existe uma métrica clara de erro a ser utilizada. Podemos apenas verificar a regularidade dos padrões (grupos) encontrados e comparar com resultados de outros estudos.

- **Aprendizado por reforço:** (*reinforcement learning*) é uma técnica de aprendizado que também é baseada em $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, mas que utiliza um sistema de recompensa e punição como sinal de resposta correta ou incorreta.

"An RL problem can be best explained through games. Let's take the game of PacMan where the goal of the agent(PacMan) is to eat the food in the grid while avoiding the ghosts on its way. In this case, the grid world is the interactive environment for the agent where it acts. Agent receives a reward for eating food and punishment if it gets killed by the ghost (loses the game). The states are the location of the agent in the grid world and the total cumulative reward is the agent winning the game."

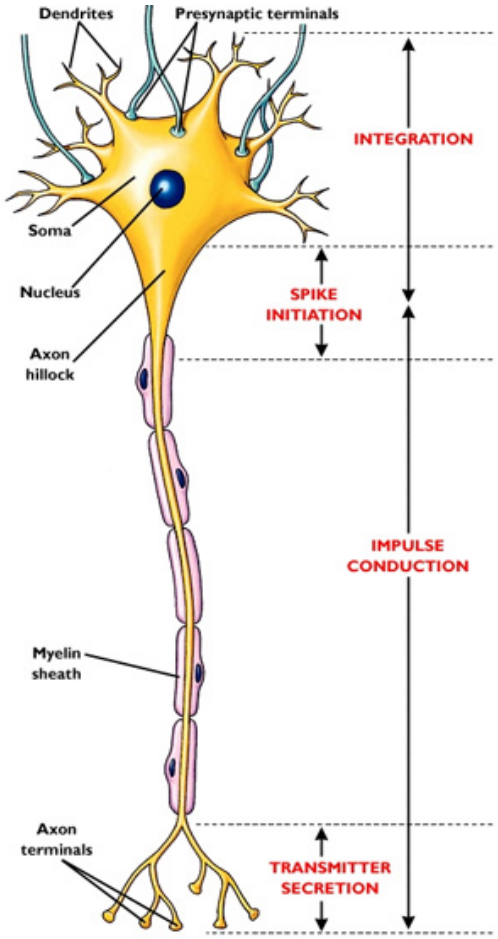


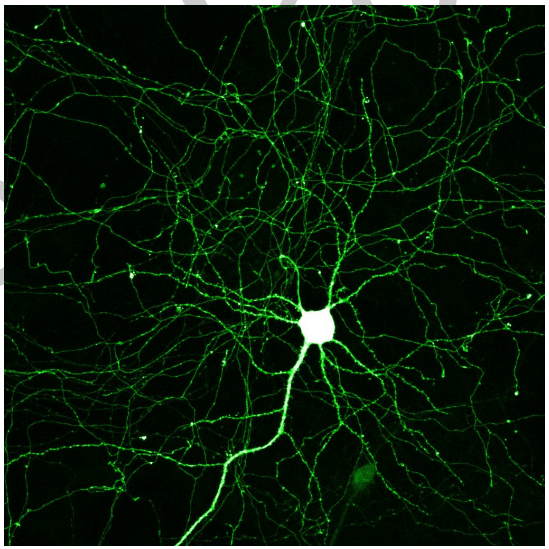
(https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292)

Aprendizado Supervisionado

- O objetivo é aprender a relação entre `inputs` **x** e `outputs` **y**, sendo $y \in \{1, \ldots, C\}$. Na classificação binária, i.e., com $C = 2$ geralmente assumimos $y \in \{0, 1\}$. Se $C > 2$ a classificação é dita politômica.

  Algumas vezes as classes não são mutuamente exclusivas (ex. carro SUV e elétrico). Neste caso denominamos de classificação multi-rótulo (*multi-label*). Este problema pode ser transformado num problema de classificação binária ou politômica?
- **Área de Aplicação:** neurociências - interface cérebro-máquina.
- **Objetivo:** classificar sinais elétricos emitidos pelo cérebro de acordo com atividade realizada.
- **Técnica de observação:** Registro de sinais elétricos do cérebro.





(impulso)

- Reportagem: Brain Power - 60 Minutes
- Experimento EST/CIC UnB - jogo pong

Abordagem Probabilística

- Em vez de uma decisão sobre a classe y_i a que pertence o objeto com vetor de características $\mathbf{x}_i$, retornamos uma medida de probabilidade de y_i pertencer a cada uma das classes possíveis.
- $\mathrm{p}(y|\mathbf{x}, \mathcal{T})$ indica a distribuição de probabilidades sobre todas as possíveis classes, dado um vetor resposta $\mathbf{x}$ e um conjunto de treinamento $\mathcal{T}$.
- Se a resposta é binária, por exemplo, $\mathrm{p}(y = 1|\mathbf{x}, \mathcal{T}) + \mathrm{p}(y = 0|\mathbf{x}, \mathcal{T}) = 1$, sendo $y = 1$ se a resposta y pertece a classe de interesse na identificação, e $y = 0$ caso contrário.
- Se vários modelos são comparados, então esta suposição deve estar explícita escrevendo $\mathrm{p}(y|\mathbf{x}, \mathcal{T}, \mathrm{M})$, em que M indica o modelo.
- Obtida uma resposta $\hat{y}$, podemos investigar o melhor "palpite" ("*guess*") procurando p que maximiza a resposta entre todas as possíveis classes C, i.e.,

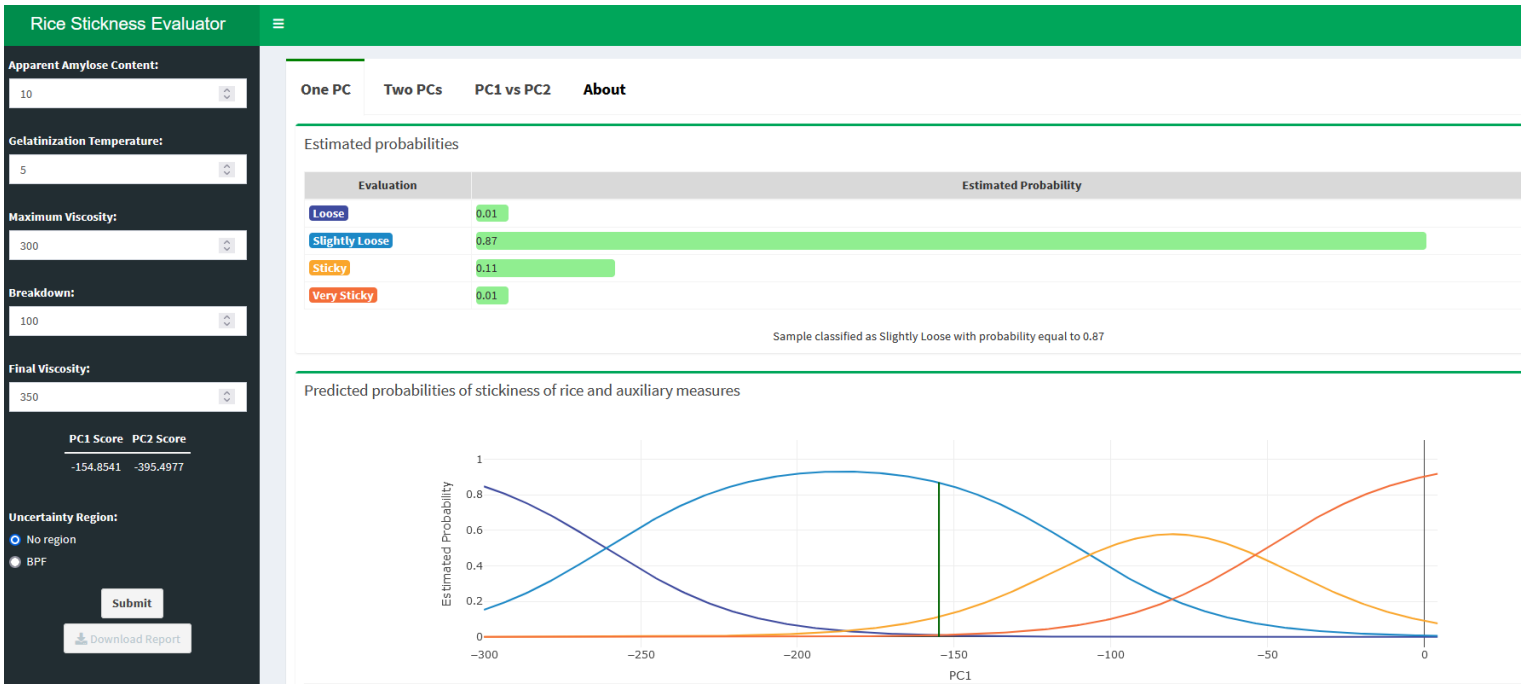
$$\hat{y} = \hat{f}(\mathbf{x}) = \arg\max_{c \in C} \ \mathrm{p}(y|\mathbf{x}, \mathcal{T}).$$

correspondendo a classe de resposta mais provável. Esta medida é também conhecida como estimativa máxima a posteriori ou MAP (*maximum a posteriori estimate*).

- Regressão Logística é um exemplo de abordagem probabilística em classificação.

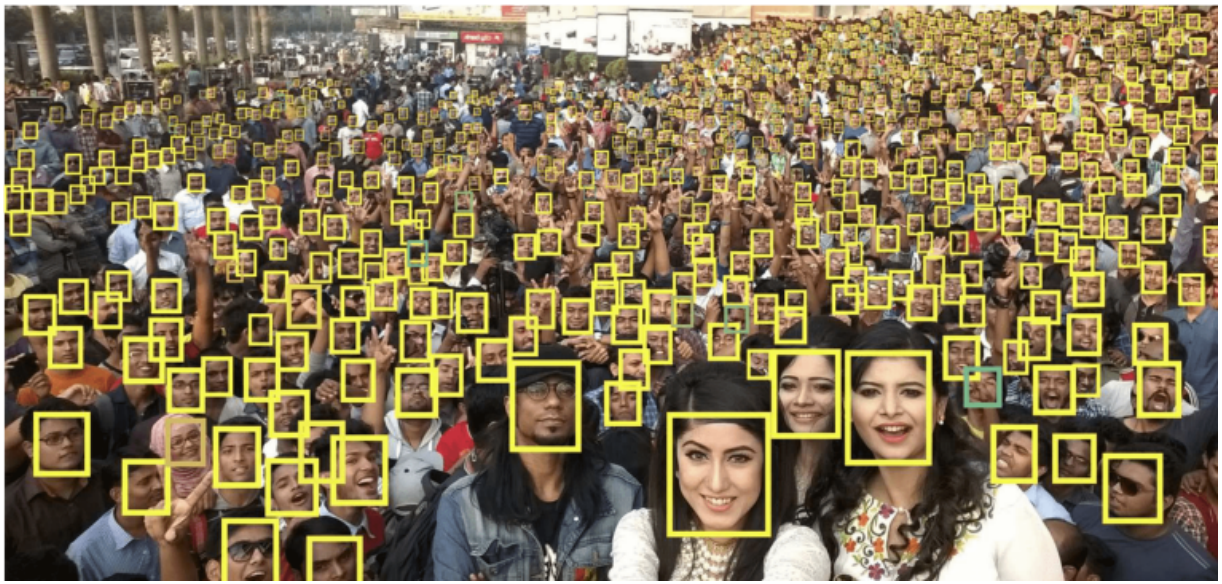
Exemplo: Viscosidade de Arroz

Avaliação de probabilidade entre 4 possíveis classes de viscosidade de arroz através de regressão logística politômica.
Ver Rice Stickness Evaluator.



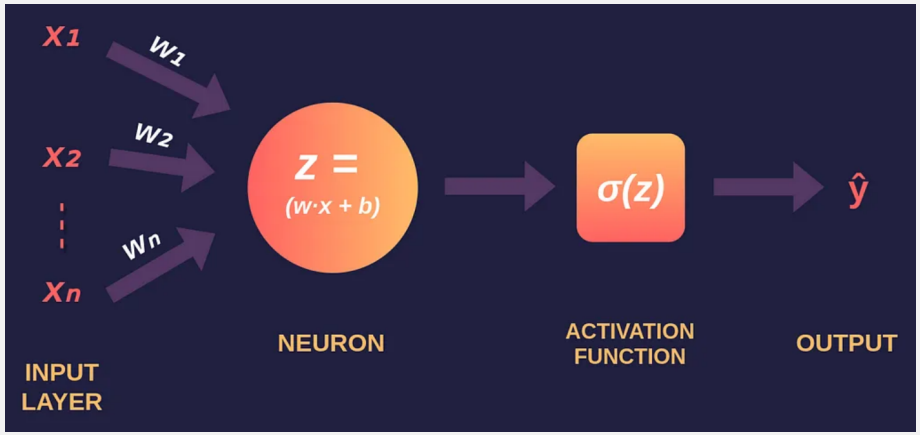
Exemplo: Detecção de Faces

Minaee, S.; Luo, P.; Lin, Z.; Bowyer, K. (2021) Going Deeper Into Face Detection: A Survey. arXiv:2103.14983v2.



Observação

Detecção de faces é um tipo de aplicação que utiliza Redes Neurais (e *Deep Learning*). Uma Rede Neural pode ser resumida na figura abaixo (Fonte: Dasaradh, 2020).



Neste curso não estudaremos Redes Neurais, mas algumas funções que resultam em z e também a ideia que envolve a função logística (ou sigmoid[a]), geralmente utilizada na função de ativação $\sigma(z)$.

[a] $\sigma(z) = \frac{1}{1+e^{-z}}$.

Comentários sobre Classificação e Agrupamentos:

1. **Extração de Características por redução de dimensão**

   A dimensionalidade do estudo pode ser restritiva para análise e processamento dos dados. Exemplos:

   - **Custo computacional:** complexidade computacional e consumo de energia.
   - **Degradação da performance:** um problema com dimensão elevada pode causar problema numéricos e convergir prematuramente para uma solução subótima.
   - **Sobreajuste dos dados:** (*overfitting*) em dados superdimensionados ($p > n$)[2] o desempenho no conjunto de treinamento pode ser muito superior ao desempenho observado no conjunto de teste.

[2] p = número de variáveise n = número de obsevações.

1. **Extração de Características por redução de dimensão**

   Procedimentos para redução de dimensão:

   - **Projeção em subespaço:** projetar os dados em um subespaço de menor dimensão, preservando a maior parte da informação. Exemplo: componentes principais, componentes independentes.

   - **Seleção de Características:** selecionar as características mais importantes através de escores que refletem o poder de discriminação (filtragem[3]) ou através da verificação do papel de cada característica na pré-classificação dos dados (*wrapper*[4]).

[3] Abordagem pré-classificação.
[4] Abordagem pós-classificação.

2 Abordagens probabilísticas para aprendizado supervisionado e não supervisionado

Estes procedimentos permitem verificar a chance de um objeto pertecer a certa classe ou se objetos tem probabilidade elevada de estarem no mesmo grupo.

- **Mistura de densidades:** a mistura de normais é a abordagem mais comum na classificação e agrupamento de dados.
- **Estimação paramétrica e não paramétrica de densidades**

Exemplo: Catalogue with probabilistic classification of galaxies based on their ionization source

Referências:

- GMM_Catalogue.
- Souza et al. (2017) A probabilistic approach to emission-line galaxy classification. arXiv:1703.07607v2.

3 Preenchimento de Matrizes (*Matrix Completion*)

Muitas variáveis do estudo podem ter dados faltantes (*missing*) por falta de resposta, falha de registro, dificuldade de coleta, etc.
A imputação na matriz procura inferir valores plausíveis e evitar a eliminação de variáveis importantes para o estudo.

Exemplo: Desafio Netflix

Competição lançada em 2006 para melhorar o sistema de recomendação de filmes para clientes. A matriz de dados tinha $n = 17.770$ filmes (colunas) e $m = 480.189$ clientes (linhas), com avaliação de alguns filmes na escala de 1 (pior) a 5 (melhor). A matriz de treinamento tinha cerca de 100 milhões de avaliações (apenas de 1% do total) e portanto com alta esparsidade.

O vencedor do desafio seria o primeiro a reduzir o RMSE do algoritmo padrão em pelo menos 10%.

- Bennet, J. e Lanning, S. (2007) The Netflix Prize. *Proceedings of the KCD Cup Workshop 2007*, New York..
- Hastie, T.; Tibshirani, R.; Wainwright, M. (2015) Statistical Learning with Sparsity (The Lasso and Generalizations). CRC Press.

Exemplo: "Market Basket Analysis"

"In commercial data mining, there is much interest in a task called market basket analysis. The data consists of a (typically very large but sparse) binary matrix, where each column represents an item or product, and each row represents a transaction. We set $x_{ij} = 1$ if item j was purchased on the i'th transaction. Many items are purchased together (e.g., bread and butter), so there will be correlations amongst the bits. Given a new partially observed bit vector, representing a subset of items that the consumer has bought, the goal is to predict which other bits are likely to turn on, representing other items the consumer might be likely to buy. (Unlike collaborative filtering, we often assume there is no missing data in the training data, since we know the past shopping behavior of each customer.)

This task arises in other domains besides modeling purchasing patterns. For example, similar techniques can be used to model dependencies between files in complex software systems. In this case, the task is to predict, given a subset of files that have been changed, which other ones need to be updated to ensure consistency.

... we can adopt a probabilistic approach, and fit a joint density model $\mathrm{p}(x_1, \ldots, x_p)$ to the bit vectors."

(Murhpy, 2012.)

4 Modelos Paramétricos e Não Paramétricos

Suponha um experimento com espaço amostral Ω. Seja $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ uma amostra aleatória com vetor definido no espaço amostral.

O resultado de um experimento é a realização $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ de um vetor aleatório $\mathbf{X}$.

Associado a $\mathbf{X}$ existe uma coleção (família) $\mathcal{F}$ de distribuições de probabilidades. $\mathcal{F}$ representa o conjunto de todas as possíveis distribuições de probabilidade para $\mathbf{X}$ e chamamos de modelo (ou modelo probabilístico) para o experimento.

Quando $\mathcal{F}$ é indexado por um ou mais parâmetros, i.e., $\mathcal{F} = \{F_\theta : \boldsymbol{\theta} \in \Theta\}$[5], Θ é chamado de espaço paramétrico em que cada F_θ é uma distribuição de probabilidade.

Desta forma, se $\mathcal{F}$ pode ser expresso como uma coleção de distribuições indexadas por vetores dimensionais finitos $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\}$, em que Θ é um subconjunto de $\Re_k$, então $\mathcal{F}$ será chamado de família de paramétrica. Se $\mathcal{F}$ não pode ser representado desta forma, então será chamado de não paramétrico.

Exemplos de Modelos

Paramétrico: estimação de densidade via modelo de mistura de normais.

Não paramétrico: estimação de densidade via método do Histograma.

[5] O vetor de parâmetros é representado por $\boldsymbol{\theta}$ e Θ representa a coleção de todos os possíveis valores de $\boldsymbol{\theta}$

5 "Maldição da Dimensionalidade" (*Curse of Dimensionality*)

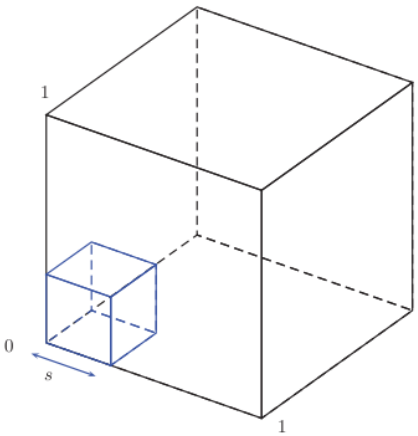
- Difícil ter dados suficientes para cobrir todas as regiões de um espaço de dimensão elevada. Isto impede aprender qual parte do espaço é importante para a relação e qual parte pode ser ignorada.

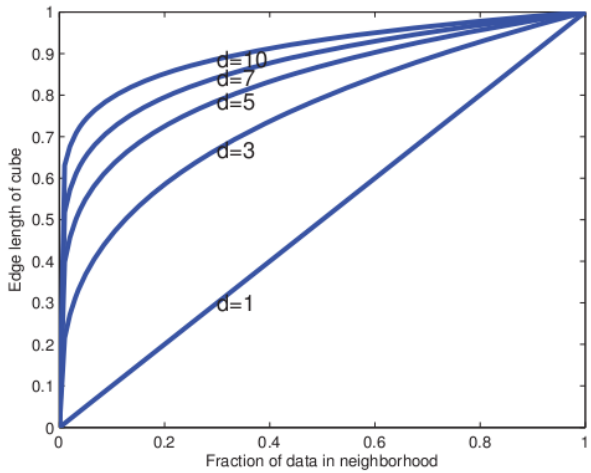
- A medida que o número de dimensões aumenta, quase toda a informação dentro da região de um hipercubo se concentra próxima da superfície do hipercubo, em vez do centro.

5 "Maldição da Dimensionalidade" (*Curse of Dimensionality*)

Um hipercubo r-dimensional $[-A, A]^r$ com vértice de comprimento $2A$ tem volume $(2A)^r$. Considere um cubo um pouco menor com vértice $2(A-\epsilon)$, $\epsilon > 0$ e pequeno.
A diferença entre o volume destes hipercubos é $(2A)^r - 2^r(A-\epsilon)^r$. Assim, a proporção do volume que esta entre os dois hipercubos é

$$\frac{(2A)^r - 2^r(A-\epsilon)^r}{(2A)^r} = 1 - \left(1 - \frac{\epsilon}{A}\right)^r \longrightarrow 1 \quad \textit{para} \quad r \longrightarrow \infty$$



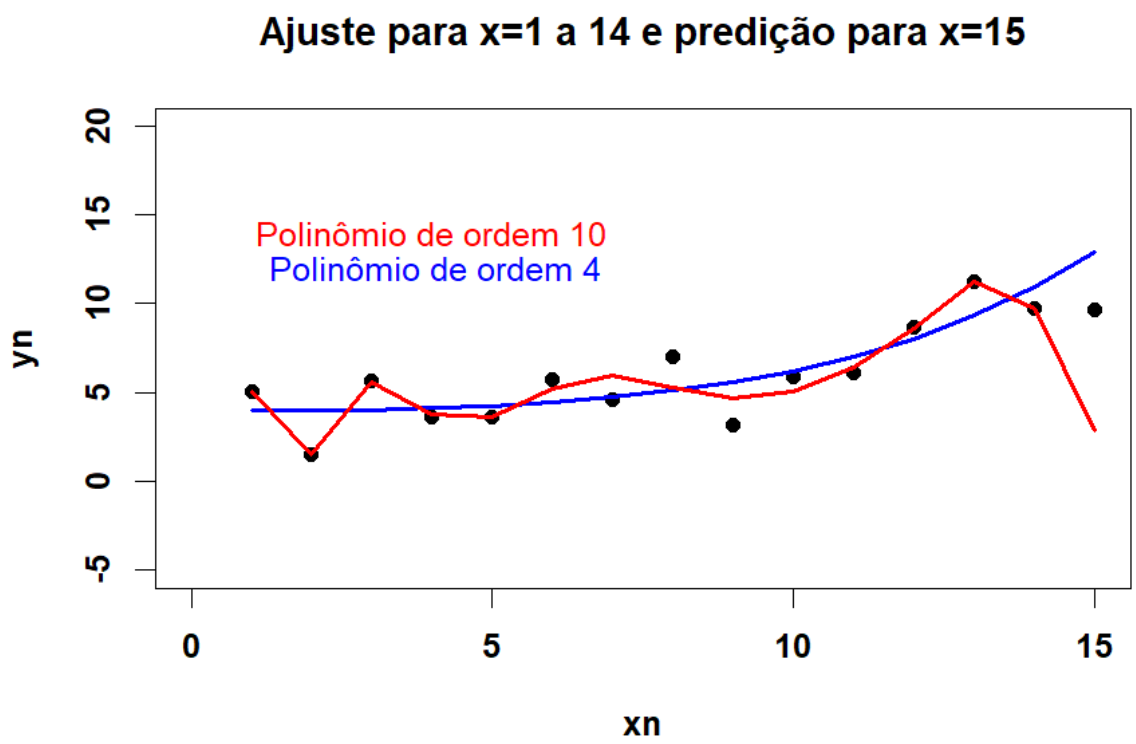


Fonte: Murphy, 2012. Machine Learning: A probabilistic perspective. MIT Press.

6 Overfitting

O ajuste de todas as pequenas variações dos dados poderá refletir ruído e não o sinal original. O erro de aprendizado (treinamento) é muito pequeno, mas o erro de generalização (teste) é elevado.

Exemplo:



"We must avoid false confidence bred from an ignorance of the probabilistic nature of the world, from a desire to see black and white where we should rightly see gray."

(em Murphy, 2022)

Etapas no Reconhecimento de Padrões

1. Obtenção dos Dados (sensor).
2. Representação do padrão a ser investigado - imagem, sinal, ...
3. Seleção/extração de características: seleção das variáveis, redução de dimensão - regularização, PCA, ...

Característica: quaisquer valores, estatísticas, padrões ou variáveis utilizadas para discriminar duas ou mais classes.

Extração de características: quaisquer técnicas ou transformações que, aplicadas a um determinado conjunto de dados, produzem um outro conjunto formado por características.

4. Treinamento da função discriminante: ajuste da função com base em alguma métrica de erro em relação a variável de supervisão. O ajuste é feito em dados de treinamento.

   Estimação estimação (modelagem) via métodos paramétricos e não paramétricos em dados de treinamento - regressão linear, regressão logística, suavização, classificação, *random forests*, máquinas de suporte vetorial, agrupamento, ...

5. Teste/Avaliação do classificador: para um conjunto de dados em separado, verificar a qualidade da função discriminante ajustada.

   Validação cruzada, bootstrap, *bagging*, *boosting*, ...
   (retornar ao item 4 sempre que necessário)

   No free lunch theorem:
   "All models are wrong, but some models are useful."

   (George Box, 1987)

6. Decisão - alocação, predição, projeção.

Referências:

1. Murphy, K. P. (2012) Machine Learning: A Probabilistic Perspective. MIT Press.
2. S. Theodoridis e K. Koutroumbas (2009) Pattern Recognition. Academic Press.
3. Hastie, T.; Tibshirani, R. e Friedman, J. (2009) The Elements of Statistical Learning (Data Mining, Inference and Prediction). Springer.