

Aprendizagem de Máquina em Física de Altas Energias

Aula 01: Reconhecimento de Padrões e Sistemas de Classificação

Prof. Edmar Egidio P. de Souza

Universidade Federal da Bahia - UFBA
Departamento de Engenharia Elétrica e de Computação
Escola INCT de Análise de Dados 2024
edmar.egidio@cern.ch

4 de novembro de 2024

Sistemas de Classificação: Detecção Binária de Sinais

- 1** Introdução a Detecção de Sinais
- 2** Detecção Binária de Sinais
- 3** Análise Discriminante
- 4** Introdução às Redes Neurais Artificiais
- 5** Considerações Finais

Introdução

- Diferentes áreas da Física de Altas Energias (HEP) tem sido exploradas com ferramentas baseadas em Aprendizado de Máquina;
- A possibilidade de exploração de correlações não-lineares entre as variáveis de entrada, aprendizado estatístico, e velocidade de processamento em execução, tem potencializado o uso de máquinas de aprendizagem em diferentes problemas em HEP.
- Identificação de eventos raros, estimativa de energia, Calibração de Calorímetros, Reconstrução de traços de partículas carregadas, Trigger e regressão de massa, são alguns exemplos.
- "*Machine Learning Everywhere*" ... Mas nem sempre foi assim.

Introdução

A área de Aprendizagem de Máquina sempre foi muito questionada em HEP:

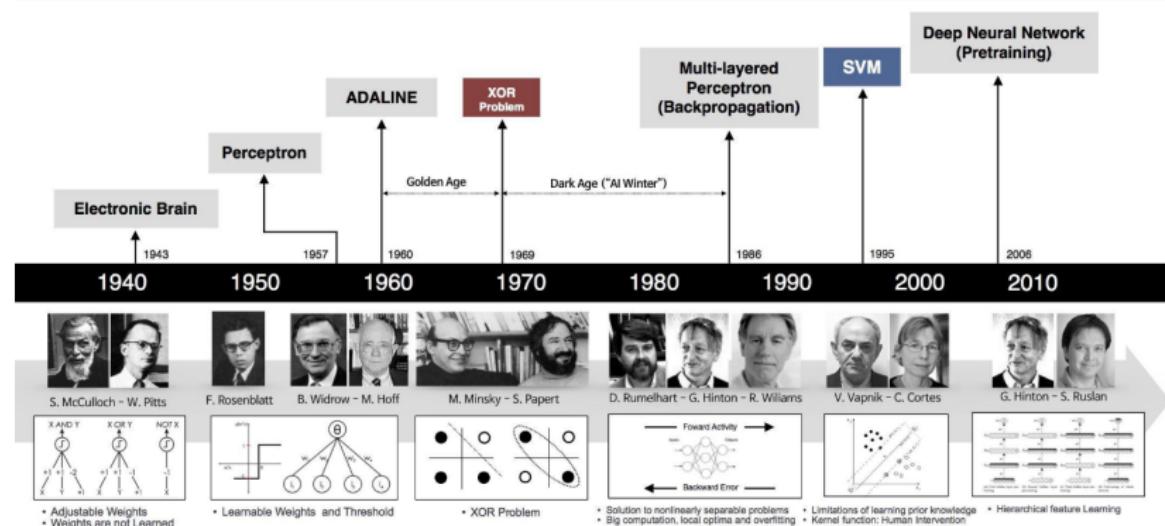
- Dificuldade de explicabilidade dos modelos;
- Estratégia de validação dos resultados com elevada dependência do contexto;
- Limitações anteriores de computação de alto desempenho;

Com a evolução das técnicas de treinamento, hardware, algoritmos de otimização e estratégias estatísticas tem sido possível utilizar Aprendizagem de Máquina para diferentes aplicações.

- O Aprendizado de Máquina guiado pelo conhecimento especialista tem sido chave para o desenvolvimento das aplicações.

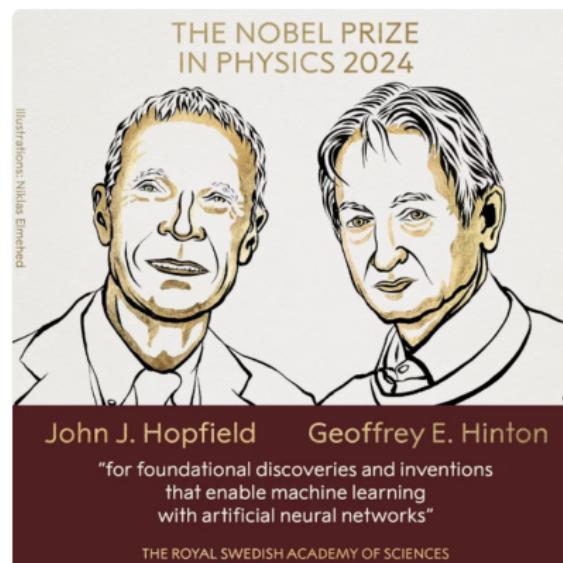
Introdução

Linha do tempo de Redes Neurais:



Introdução

Nobel de Física 2024 vai para dois cientistas com estudos sobre aprendizado de máquina:



Introdução

Em Física de Altas Energias, a área de reconhecimento de padrões encontra aplicações distintas, como:

- Classificação Sinal/Ruído;
- Sistemas de Filtragem e Identificação Online de Partículas; (*Trigger*)
- Calibração e Regressão de Energia;
- Estimação de diferenças dado/monte carlo;
- Reconstrução de trajetória de partículas carregadas;

MAGIC Telescope - Separação γ /hadron



The Trigger System of the MAGIC Telescope

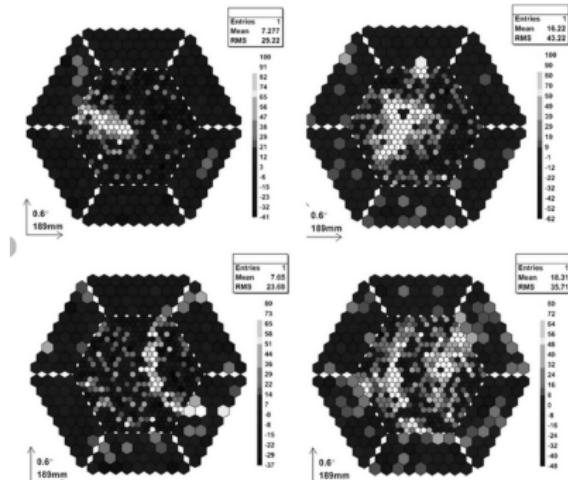
R. Paoletti, R. Cecchi, D. Corti, F. Dazzi, M. Mariotti, R. Pegna, and N. Turini

Abstract—The MAGIC telescope aims at the detection of very low energy gamma rays ($E > 30$ GeV) through the atmospheric emission of Cherenkov light. The high background rate originating from night sky background, atmospheric showers and bright stars sets a serious challenge to this goal. Application of topological selection cuts at trigger level can have a big impact on background reduction allowing the telescope to operate at the threshold of the available background. The trigger of the MAGIC telescope is a two-level system following a pipeline philosophy, similar to those adopted in high energy physics experiments. Operative since October 2002, the trigger system has been a key point in the commissioning of the MAGIC telescope and is now taking data. The trigger hardware is described in detail.

Index Terms—Astroparticle physics, astrophysics, imaging air Cherenkov telescope, trigger.



Fig. 1. The MAGIC telescope. On the left is the access tower, used to work on the cameras.



Typical shower images in the MAGIC Telescope, corresponding to different candidate events: gamma (top-left), hadron (top-right), muon (bottom-left) and muon plus hadron (bottom-right).

MAGIC Telescope - Separação γ /hadron

The performance of the MAGIC telescopes using deep convolutional neural networks with CTLearn

T. Miener,^{a,*} D. Nieto,^a R. López-Coto,^b J. L. Contreras,^a J. G. Green,^c D. Green^c and E. Mariotti^d on behalf of the MAGIC Collaboration

^aInstituto de Física de Partículas y del Cosmos y Departamento de EMFTEL, Universidad Complutense de Madrid, Spain

^bInstituto de Astrofísica de Andalucía - CSIC, Granada, Spain

^cMax-Planck-Institut für Physik, München, Germany

^dDipartimento di Fisica e Astronomia dell'Università e Sezione INFN, Padova, Italy

E-mail: tmienner@ucm.es

Observational data or simulations

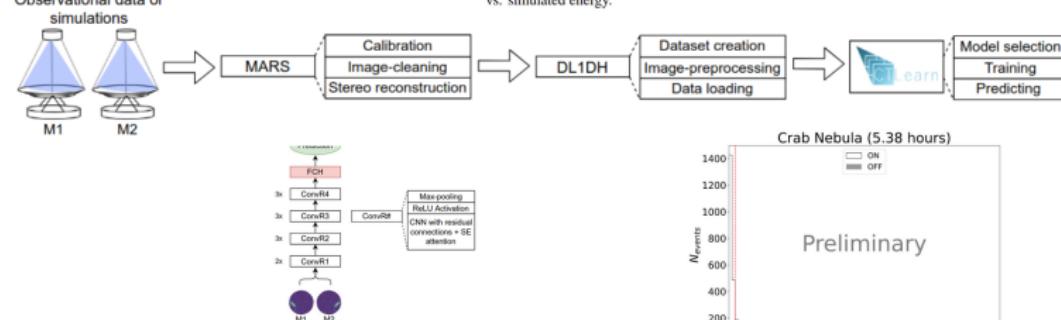


Figure 2: CTLearn's TRN model with channel-wise concatenation of the two stereoscopic images recorded by the MAGIC telescopes (M1 and M2).

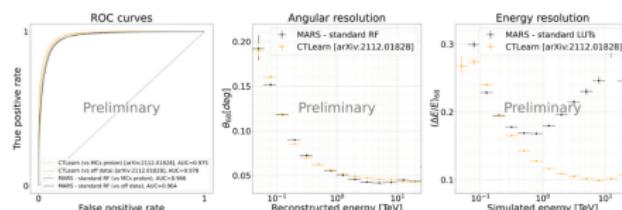
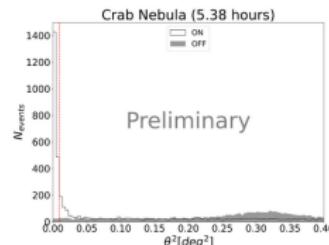


Figure 3: The validation of the performance is taken from [8]. Left) ROC curves with MC proton simulations and observational off data. Center) Angular resolution vs. reconstructed energy. Right) Energy resolution vs. simulated energy.



Seleção *Online* de Eventos no ATLAS/LHC

Porque um sistema *online* de seleção de eventos no ATLAS?

- Frequência de colisões em 40 MHz. Um evento de colisão apresenta $\approx 1,7$ MB. A taxa de saída de dados total (sem *trigger*) seria de ≈ 70 TB/s;
- Os eventos físicos de interesse são raros, em um ambiente com intenso ruído de fundo;
- Nem tudo pode ser armazenado: Deve ser eficiente na física de interesse e ainda rejeitar a maioria dos eventos não relevantes.
- O que não é selecionado pelo *trigger* \rightarrow ***LOST FOREVER***.

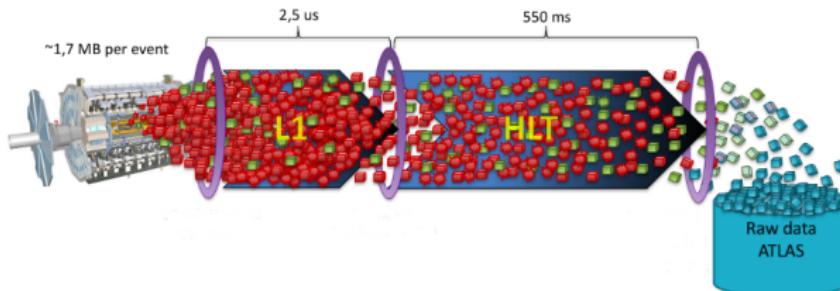


Ilustração do sistema *online* de seleção de eventos empregado no ATLAS.

Algoritmo NeuralRinger - Detecção de Online de Elétrons

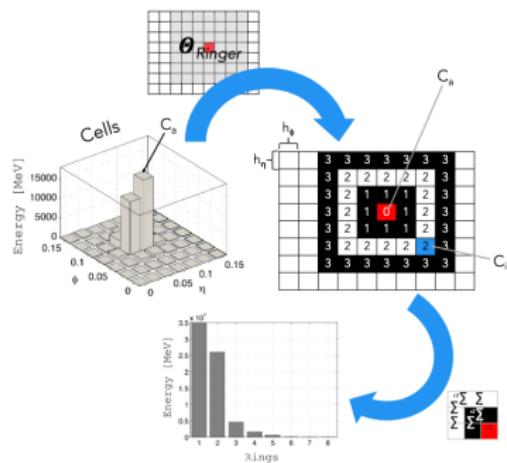


Ilustração da extração dos anéis a partir da informação das células.

- O algoritmo NeuralRinger atua na etapa rápida do HLT.
- Captura a informação lateral e longitudinal do chuveiro de partículas em uma geometria aproximadamente cônica;
- Descreve o perfil de deposição de energia da partícula em anéis concêntricos em torno da célula mais quente;
- Um total de 100 anéis são gerados, distribuídos em cada camada do sistema de calorimetria;
- Um conjunto de classificadores neurais, recebe como entrada os sinais em anéis e realiza a decisão e^-/j ;

Introdução

Definição

O reconhecimento de padrão é formalmente definido como o processo pelo qual um padrão/sinal recebido é atribuído a um número prescrito de classes (Haykin, 2020).

Introdução

Definição

O reconhecimento de padrão é formalmente definido como o processo pelo qual um padrão/sinal recebido é atribuído a um número prescrito de classes (Haykin, 2020).

- O projeto de máquinas de aprendizagem para reconhecimento de padrões, devem ser capazes de fornecer uma resposta razoável para uma dada entrada e realizar a correspondência "mais provável" de um padrão;

Introdução

Definição

O reconhecimento de padrão é formalmente definido como o processo pelo qual um padrão/sinal recebido é atribuído a um número prescrito de classes (Haykin, 2020).

- O projeto de máquinas de aprendizagem para reconhecimento de padrões, devem ser capazes de fornecer uma resposta razoável para uma dada entrada e realizar a correspondência "mais provável" de um padrão;
- Necessário considerar a variação estatística dos atributos de entrada das distintas classes.

Introdução

Reconhecimento de Padrões

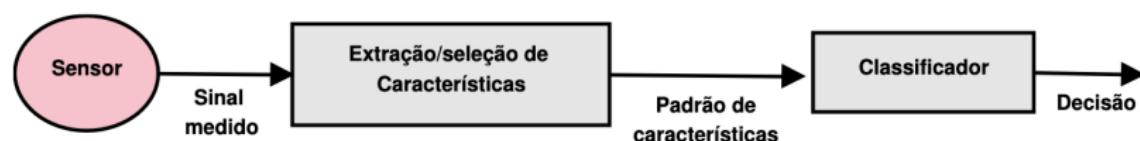
A detecção (ou classificação) de sinais consiste em identificar, a partir de observações sujeitas a variações aleatórias, o tipo (ou classe) de sinal que foi originalmente enviado.

- No projeto de um sistema de detecção pode-se citar dois parâmetros muito importantes: Eficiência de discriminação (maximizar acertos e minimizar erros) e tempo de execução da cadeia de processamento de sinais.
- Sistemas de classificação podem ser projetados para a detecção de duas classes (detecção binária), ou em problemas com múltiplas classes (multi-classes).

Introdução

Um sistema completo de reconhecimento de padrões em geral pode ser composto por:

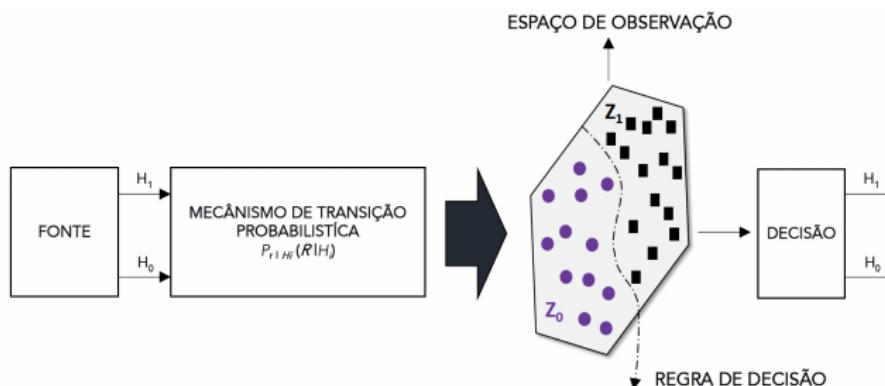
- Sensores que obtém observações a serem classificadas ou descritas;
- Um mecanismo de extração de características que computa informações numéricas ou simbólicas das observações;
- E um esquema de classificação das observações, que depende das características extraídas.



Decisão Binaria

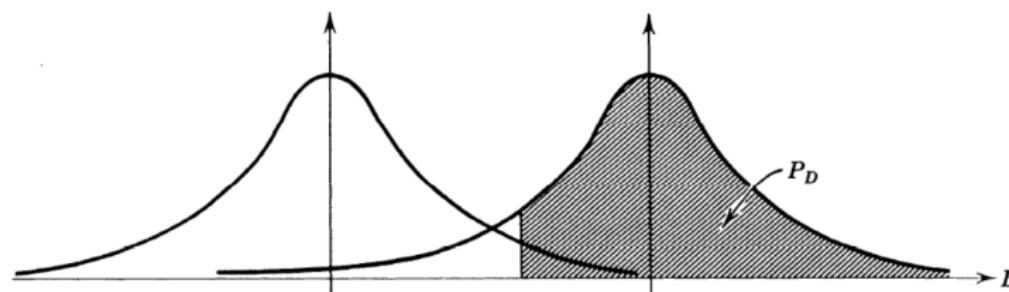
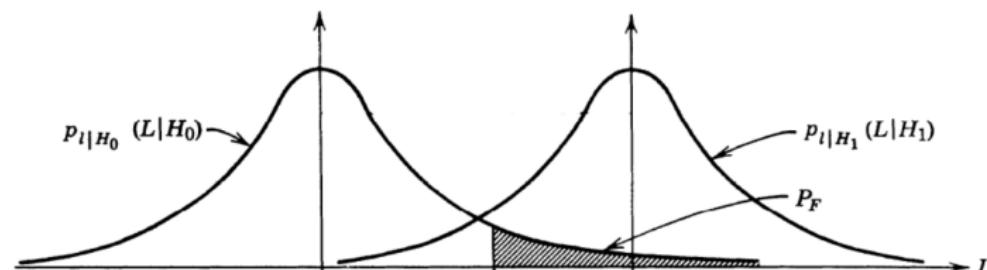
Um sistema de classificação binário corresponde a um modelo que realiza a escolha entre duas hipóteses:

- H_1 , normalmente associada à ocorrência do evento de interesse.
- H_0 , que representa a ocorrência de evento não relevante.



Decisão Binária

Considerando que a decisão é baseada na observação da variável aleatória L :



Decisão Binária

O resultado da tarefa de classificação binária pode gerar quatro situações distintas:

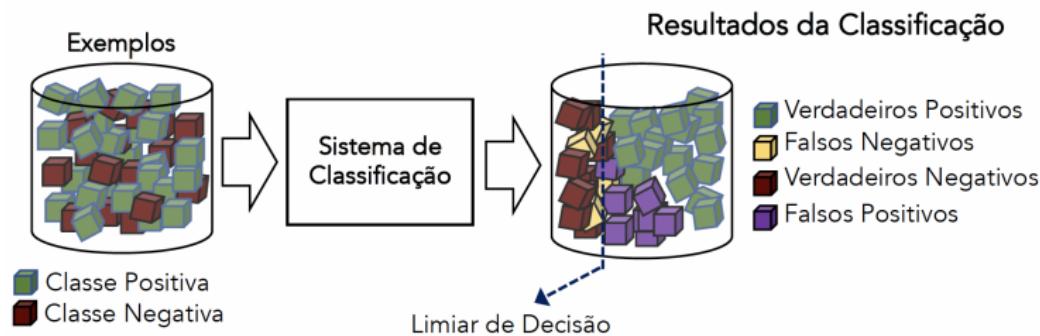


Figura: Ilustração que exemplifica os possíveis resultados de um sistema binário de classificação.

Decisão Binária

Verdadeiros Positivos (VP): Exemplos da classe positiva classificados corretamente:

$$TVP = \frac{VP}{VP + FN} \quad (1)$$

Decisão Binária

Verdadeiros Positivos (VP): Exemplos da classe positiva classificados corretamente:

$$TVP = \frac{VP}{VP + FN} \quad (1)$$

Verdadeiros Negativos (VN): Exemplos da classe negativa classificados corretamente:

$$TVN = \frac{VN}{VN + FP} \quad (2)$$

Decisão Binária

Verdadeiros Positivos (VP): Exemplos da classe positiva classificados corretamente:

$$TVP = \frac{VP}{VP + FN} \quad (1)$$

Verdadeiros Negativos (VN): Exemplos da classe negativa classificados corretamente:

$$TVN = \frac{VN}{VN + FP} \quad (2)$$

Falsos Positivos (FP): Eventos pertencentes a classe negativa, classificados incorretamente como classe positiva.

$$TFP = \frac{FP}{FP + VN} \quad (3)$$

Decisão Binária

Verdadeiros Positivos (VP): Exemplos da classe positiva classificados corretamente:

$$TVP = \frac{VP}{VP + FN} \quad (1)$$

Verdadeiros Negativos (VN): Exemplos da classe negativa classificados corretamente:

$$TVN = \frac{VN}{VN + FP} \quad (2)$$

Falsos Positivos (FP): Eventos pertencentes a classe negativa, classificados incorretamente como classe positiva.

$$TFP = \frac{FP}{FP + VN} \quad (3)$$

Falsos Negativos (FN): Exemplos pertencentes a classe positiva, classificados incorretamente como classe negativa:

$$TFN = \frac{FN}{VP + FN} \quad (4)$$

Curva ROC

O patamar de decisão do sistema de classificação pode ser escolhido através da análise da curva ROC (*Receiver Operating Characteristics*):

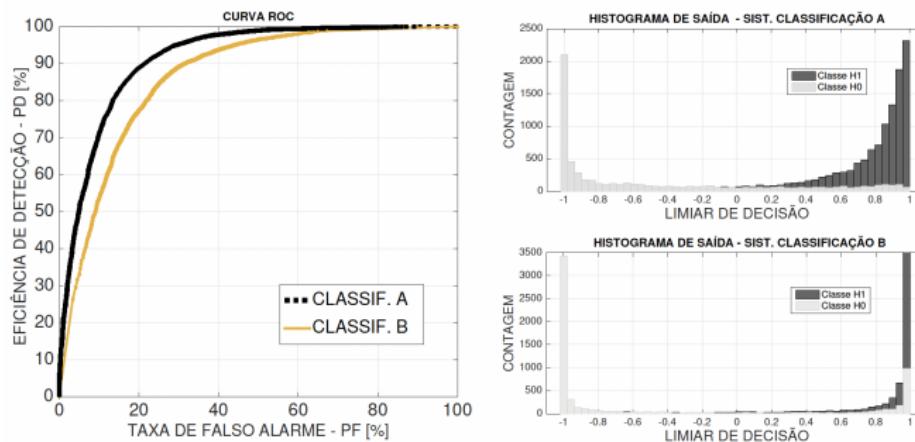
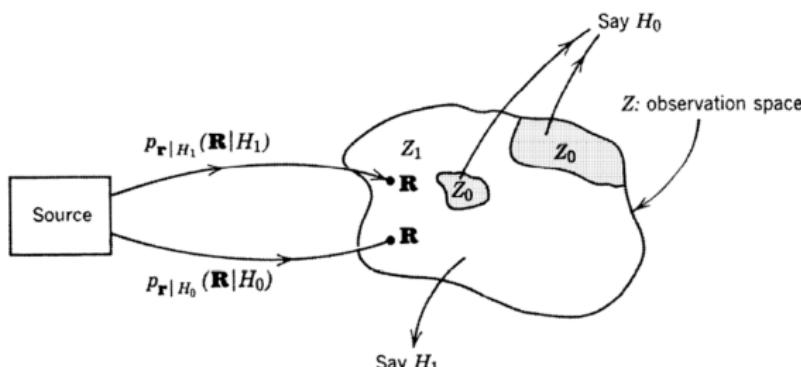


Figura: Curvas ROC de dois classificadores binários hipotéticos distintos e os respectivos histogramas de saídas

Decisão Binaria Baseada em uma Única Observação

Considerando que cada uma das hipóteses é associada a uma saída que é mapeada numa região do espaço de observação de dimensão N, um ponto neste espaço pode ser representado pelo vetor r .

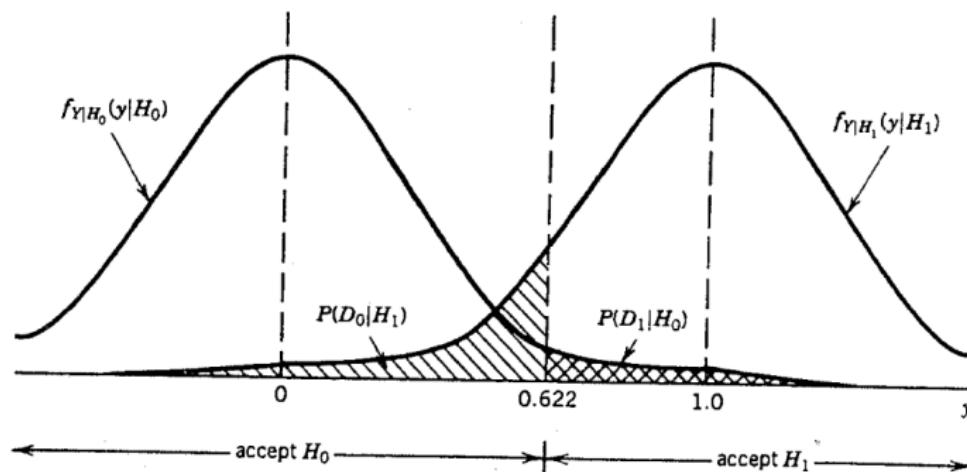
- Uma regra de decisão (patamar) produz o particionamento do espaço de observação em duas regiões Z_1 e Z_0 , para as quais são associadas as hipóteses H_1 e H_0 , dependendo da observação de r .



Definição do Patamar de Decisão entre Classes

- A probabilidade condicional $P_{r/H_i}(R/H_i)$ é chamada de probabilidade a posteriori, pois são estimadas apos a observação de r .
- Quando as densidades de probabilidade condicionais são conhecidas (ou podem ser estimadas), o projeto do sistema de reconhecimento de padrões pode ser simplificado.
- Os critérios de **Máximo a Posteriori**, **Bayes**, **Minimax** e **Neyman-Pearson** são procedimentos clássicos utilizados para a escolha da regra de decisão quando as probabilidades condicionais são conhecidas.

Exercício: Critério MAP



Analise de Discriminantes

Embora muito utilizadas pela formulação matemática relativamente simples e bom desempenho em diversas aplicações, as técnicas discutidas necessitam de conhecimento prévio a respeito:

- Das distribuições de probabilidade;
- Dos custos associados as classes;

Em muitos casos práticos essas informações não estão disponíveis, sendo necessário a utilização de outros métodos de classificação.

- Uma opção neste contexto é utilizar a analise de discriminantes.

Analise de Discriminantes

- Os métodos de classificação baseados na regra de Bayes, particionam o espaço de observações em regiões associadas a cada hipótese.
- A analise de discriminantes , de modo análogo, busca as superfícies que particionam o espaço de observações de modo "ótimo" nas regiões que são associadas a cada hipótese.
- Uma função discriminante linear pode ser dada por:

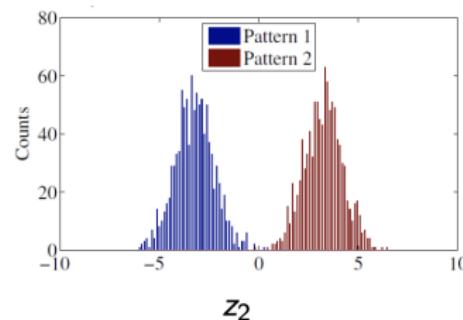
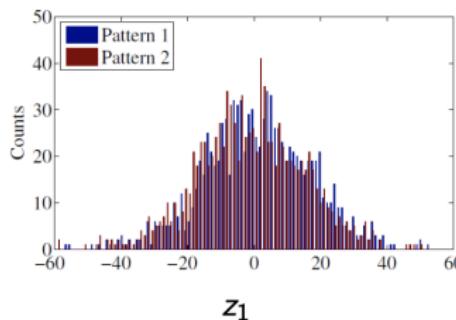
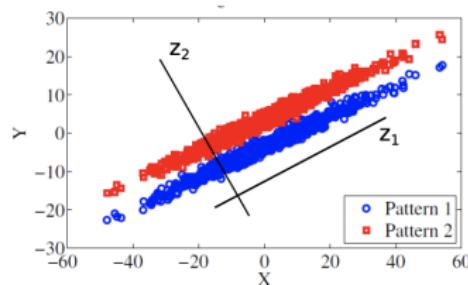
$$y(x) = \mathbf{w}^T \mathbf{x} + \omega_0, \quad (5)$$

Sendo \mathbf{w} o vetor de pesos e ω_0 a tendência (ou bias).

- A fronteira de decisão é definida por $y(x) = 0$

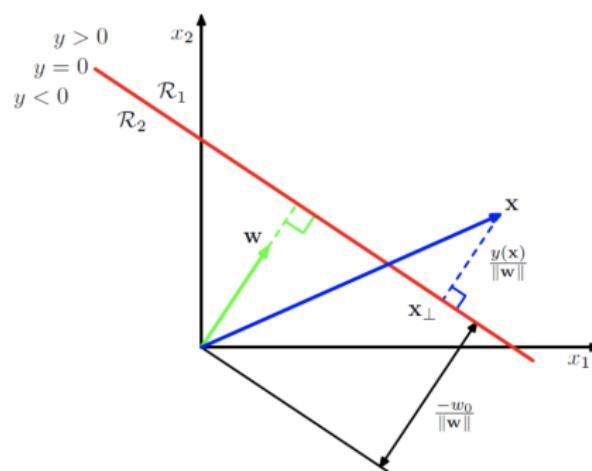
Analise de Discriminantes

- A análise de discriminantes busca a direção w onde as projeções $y(x)$ dos sinais de entrada x sejam maximamente separáveis.



Analise de Discriminantes

- \mathbf{w} é ortogonal à superfície de separação.
- A distância entre a superfície de separação e a origem é dada por: $\frac{-\omega_0}{\|\mathbf{w}\|}$.
- A distância entre um ponto \mathbf{x} e a superfície de separação é: $\frac{y(\mathbf{x})}{\|\mathbf{w}\|}$.
- Como estimar a superfície de separação “ótima” a partir dos dados disponíveis?



Analise de Discriminantes - Discriminante Linear de Fisher

- A análise por discriminante de Fisher (FDA - *Fisher Discriminant Analysis*) busca a direção ótima de discriminação utilizando 2 parâmetros, a distância inter-classes, e a distância intra-classes.
- Numa formulação matricial o objetivo é encontrar a direção \mathbf{w}_0 que maximiza a expressão:

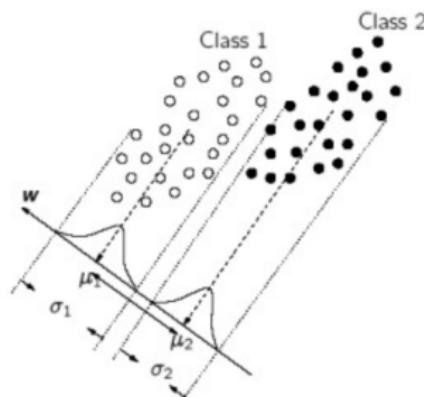
$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$$

onde $\mathbf{S}_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ é a matriz de separação inter-classes e $\mathbf{S}_w = \mathbf{S}_1 + \mathbf{S}_2$ é a matriz de separação intra-classes, sendo:

$$\mathbf{S}_i = \sum_{\mathbf{x} \in \mathcal{D}_j} (\mathbf{x} - \mu_i)(\mathbf{x} - \mu_i)^T$$

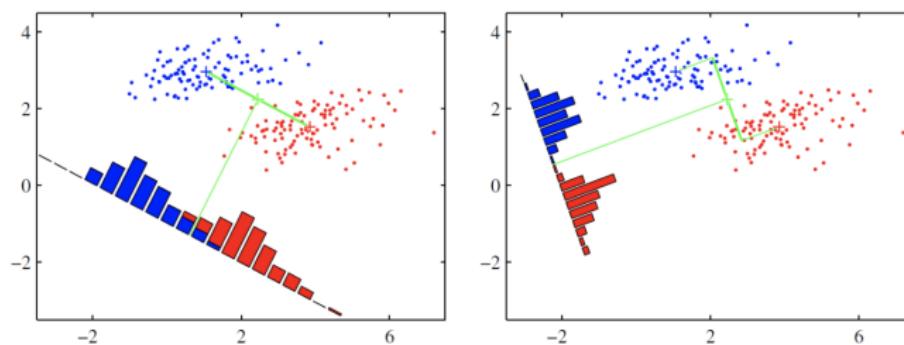
Analise de Discriminantes - Discriminante Linear de Fisher

- O discriminante de Fisher é capaz de encontrar a transformação linear ótima dos sinais de entrada, de modo que os sinais projetados $\mathbf{y} = \mathbf{w}^T \mathbf{x}$ tenham máxima separação:



Analise de Discriminantes - Discriminante Linear de Fisher

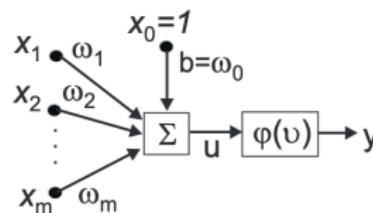
- Pode-se realizar a análise por discriminante de Fisher de modo analítico usando as equações definidas anteriormente.



- Limitações podem surgir quando a dimensão de x cresce, pois o cálculo de S_w^{-1} pode se tornar custoso computacionalmente.
- Uma opção é realizar o cálculo estimado de modo iterativo a partir de um *perceptron* (modelo básico de rede neural artificial).

Perceptron

- Outro exemplo de discriminante linear pode ser obtido através do *perceptron*.
- O modelo do *perceptron* foi proposto por Rosenblat em 1962 e foi inspirado no funcionamento de um neurônio biológico:



$$y(\mathbf{x}) = \varphi(\mathbf{w}^T \mathbf{x})$$

sendo $\mathbf{w} = [\omega_0, \omega_1, \dots, \omega_m]$ o vetor de pesos sinápticos e φ a função de ativação (normalmente é utilizada a função degrau: $\varphi(a) = 1$ para $a \geq 0$ e -1 para $a < 0$).

Perceptron

- Para o processo de treinamento, o propósito é minimizar o erro quadrático médio da classificação (sendo então baseado no algoritmo LMS - *Least Mean Square*).
- Deste modo pode-se chegar à regra de aprendizado do *perceptron*:

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \eta \mathbf{x}(n) e(n)$$

sendo:

- $e(n) = d(n) - \mathbf{w}^T(n) \mathbf{x}(n)$ o erro em relação à saída desejada $d(n)$;
- η a taxa de aprendizado.

Perceptron

- Típicas curvas de erro no treinamento de um *perceptron*:



- Espera-se que com o decorrer do treinamento o erro diminua até chegar ao seu mínimo.

Redes Neurais Artificiais

- As redes de múltiplas camadas alimentadas adiante (*feed-forward*) são compostas a partir da conexão sequencial de duas ou mais camadas de neurônios.
- Essas redes são usualmente chamadas de perceptrons de múltiplas camadas (MLP-Multi-layer Perceptrons) por serem uma generalização do perceptron.
- A saída de uma camada é utilizada como entrada da próxima. Por não possuírem laços de realimentação (redes alimentadas adiante, ou *feed-forward*) são estruturalmente estáveis.
- A camada oculta é responsável, em um processo de reconhecimento de padrões, por extrair características estatísticas de ordem superior, aplicando uma transformação não-linear aos dados de entrada.

Redes Neurais Artificiais

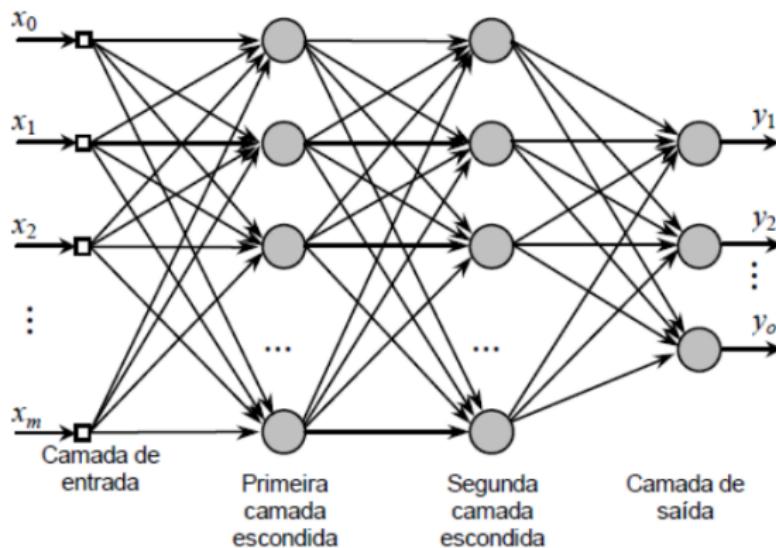


Figura: Exemplo de uma rede MLP com duas camadas intermediárias (ou escondidas, ocultas).

Treinamento de Redes Neurais Artificiais

Algoritmo de Retropropagação do Erro:

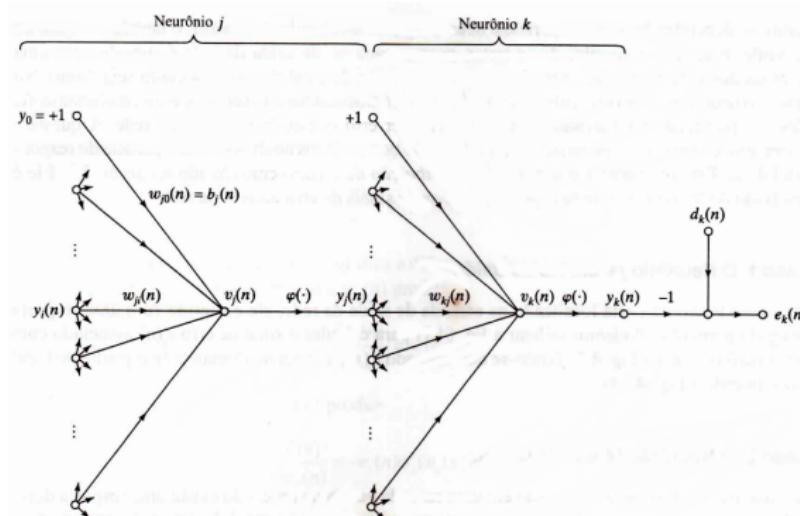


Figura: Propagação do sinal para frente (feed-forward).

Treinamento de Redes Neurais Artificiais

Algoritmo de Retropropagação do Erro:

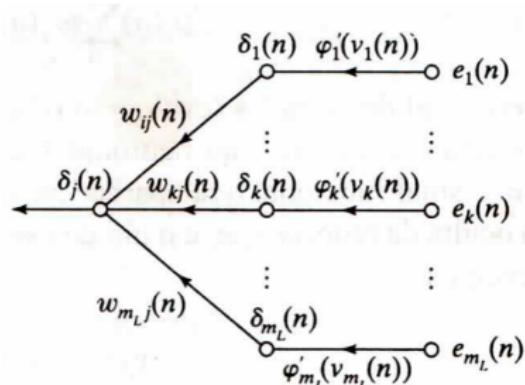


Figura: Retropropagação do erro (backpropagation).

Treinamento de Redes Neurais Artificiais

Na aprendizagem supervisionada, o modelo codifica em suas variáveis internas o conhecimento do padrão que melhor descreve as classes com base em um conjunto estatisticamente representativo com amostras rotuladas:

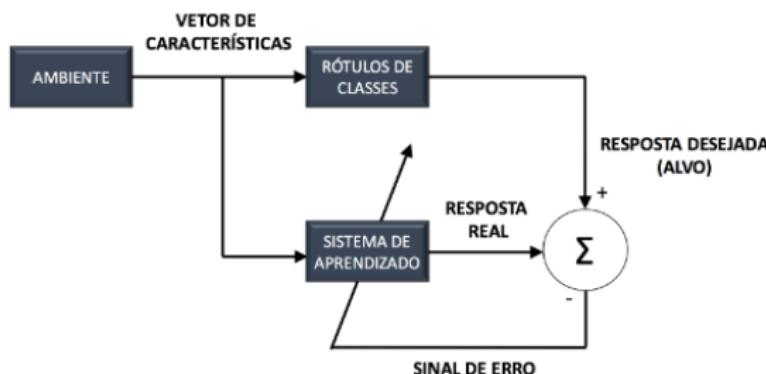


Figura: Diagrama em blocos do método de aprendizagem supervisionada para sistemas de classificação.

Treinamento de Redes Neurais Artificiais

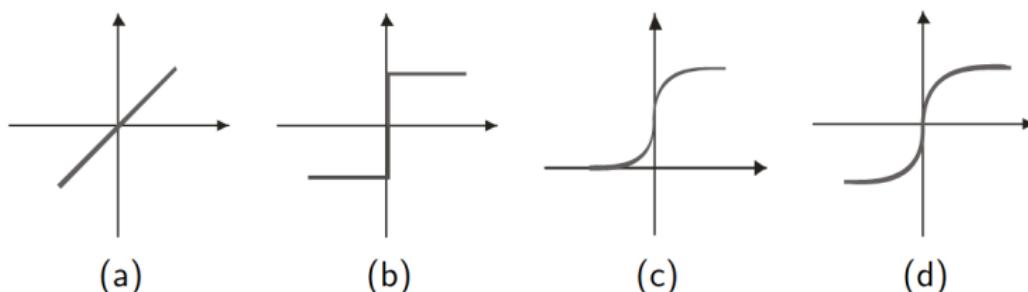
Para o projeto do classificador neural, em geral, dividem-se os pares entrada-saída disponíveis nos conjuntos de:

- Treino: utilizado para o ajuste dos pesos sinápticos.
- Validação: é na parada do treinamento prevenindo a ocorrência de sobre-aprendizado.
- Teste: utilizado para avaliar o resultado obtido e verificar a generalização do modelo, já que este conjunto não foi usado no ajuste dos pesos e ou na parada do algoritmo.

Arquitetura de Redes Neurais e Parâmetros de Ajustes

Funções de Ativação: Deve ser estudada cuidadosamente em função do problema, e do método de otimização selecionado.

- Exemplos de funções de ativação $\varphi(\cdot)$: (a) linear; (b) degrau; (c) sigmóide e (d) tangente hiperbólica.



Arquitetura de Redes Neurais e Parâmetros de Ajustes

Configuração de Hiperparâmetros:

- Épocas de Treinamento: O número de épocas é o número total de vezes que a rede neural passa por todo o conjunto de treinamento durante o treinamento. Cada época consiste em uma passagem pelo conjunto de treinamento completo.
- Critério de Parada: Critério adotado para interromper o treinamento. Ex.: número máximo de épocas, insucessos de melhoria na função objetivo, etc.
- Algoritmo de Otimização: Algoritmo iterativo selecionado para realizar o procedimento de otimização dos valores dos pesos da rede. Ex.: Newton, quasi-Newton, Gradiente Conjugado, ADAM, etc.
- Batch Size: O tamanho do lote geralmente corresponde ao número de padrões que são mostrados a rede antes dos pesos serem atualizados.
- Número de Neurônios e Camadas Ocultas: Deve ser selecionado cuidadosamente, avaliando o desempenho da rede em termos de sobreajuste, custo computacional, e convergência do algoritmo de otimização.
- Função Objetivo (loss): Medida de avaliação de desempenho utilizado para otimizar os pesos da rede.

Exercício Computacional 01

- 1 Explorar o *python notebook* com o exemplo de sistemas neurais para classificação, usando dados do MAGIC Telescope.
- 2 Avaliar o impacto da mudança de hiperparâmetros, como: Número de Épocas, Critério de Parada, Número de Neurônios Ocultos e Algoritmo de Otimização e Função Loss.

Referências

- HAYKIN, S., Neural Networks: Principles and Practice. v. 3. Prentice Hall, 2008.
- KUNCHEVA, L. I., Combining Pattern Classifiers: Methods and Algorithms. v. 2. John Wiley and Sons, 2014.
- DA FONSECA PINTO, J. V., Ring-shaped Calorimetry Information for a Neural EGamma Identification with ATLAS Detector , Tech. rep., CERN, Geneva, Mar 2016.
- BETHAPUDI, S., DESAI, S., "Separation of pulsar signals from noise using supervised machine learning algorithms", Astronomy and Computing, v. 23, pp. 15–26, 2018.
- Notas de Aula: Prof. Eduardo F. Simas Filho, Disciplina: Inteligência Computacional (ENGA74), Programa de Pós-Graduação em Engenharia Elétrica - PPGEE/UFBA.