



Supervised Learning Models on PPG Analysis

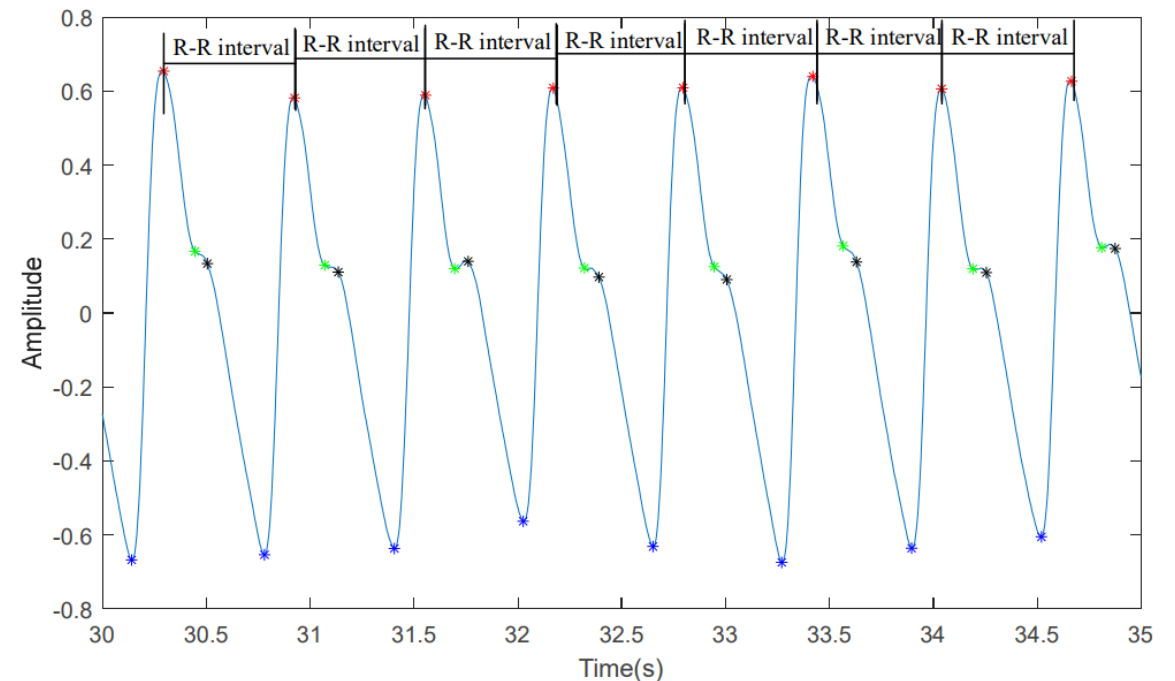
GONÇALO IVO E ARIANA OLIVEIRA

ORIENTADOR: DR^a TÂNIA PEREIRA

Importância de um sinal de pressão arterial

Onda de pressão arterial (APW) é um sinal biomédico de alta relevância. Este permite estudar diversos aspectos relativos ao sistema cardiovascular. Informações sobre a interação entre a ejeção do ventrículo esquerdo e as propriedades físicas da circulação arterial podem ser determinadas pela análise descritiva e quantitativa destes sinais. Alguns dos usos deste sinal são:

- Eficiência de ejeção de sangue e contratilidade do músculo cardíaco
- Estudo da rigidez arterial
- Estudo do envelhecimento vascular
- Determinação se a dose de anestesia administrada é a indicada
- Diagnóstico de doenças cardiovasculares como hipertensão, insuficiência cardíaca e outras doenças arteriais

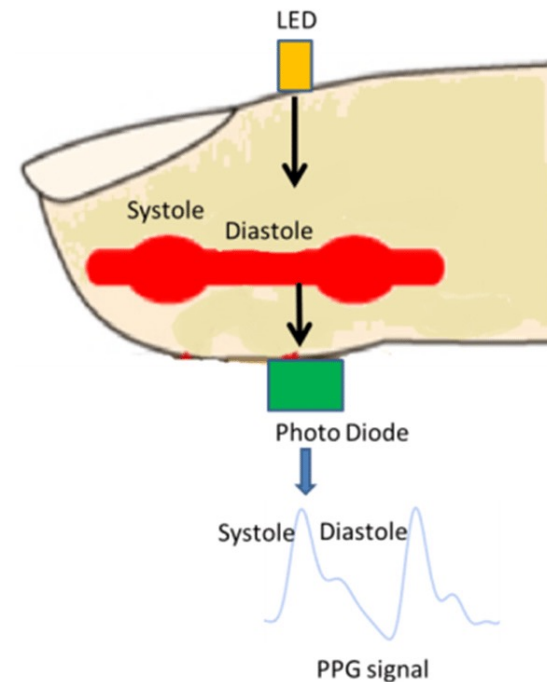


Importância de um sinal PPG

Photoplethysmography (PPG) é um destes sinais de pressão arterial, obtido através de interação ótica, usado para detetar variações no volume de sangue num determinado local.

O método mais comum é usar oximetria de pulso, um método não invasivo para medir a saturação de oxigénio no sangue. Um sensor emite luz através de um tecido (normalmente a pele do dedo) e mede a quantidade de luz absorvida pelo sangue. A variação na absorção de luz está relacionada com as variações no volume sanguíneo, que ocorrem com cada pulsação.

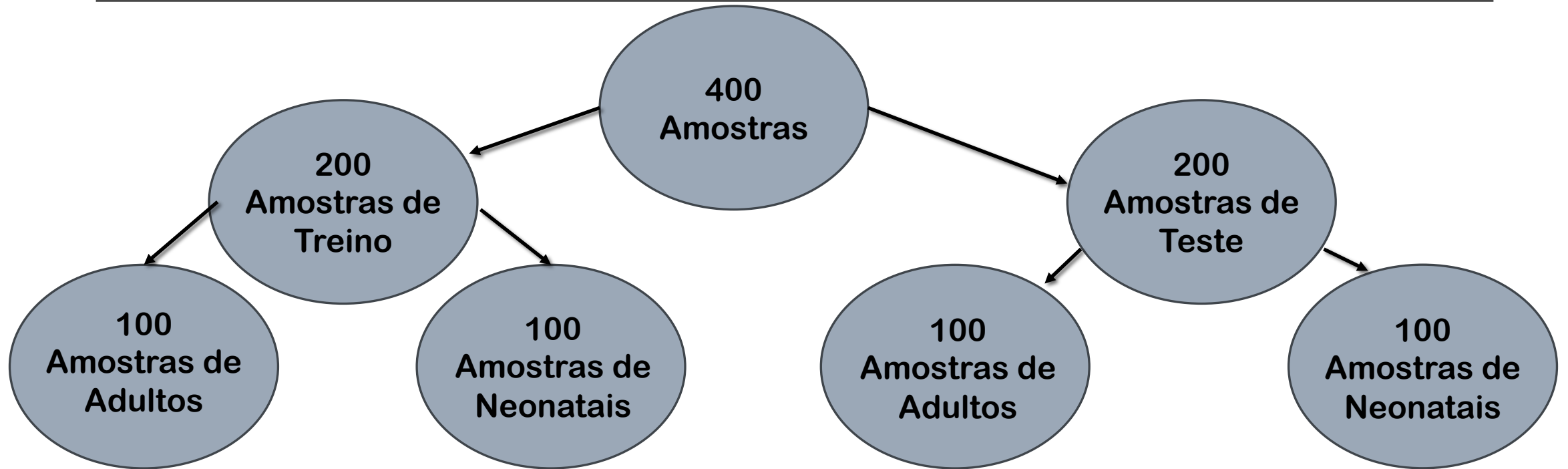
Esta técnica é aplicada em ambientes clínicos para monitorar pacientes durante cirurgias, em unidades de terapia intensiva, ou mesmo em casa para monitoramento de condições médicas específicas.



Sinal PPG Utilizado

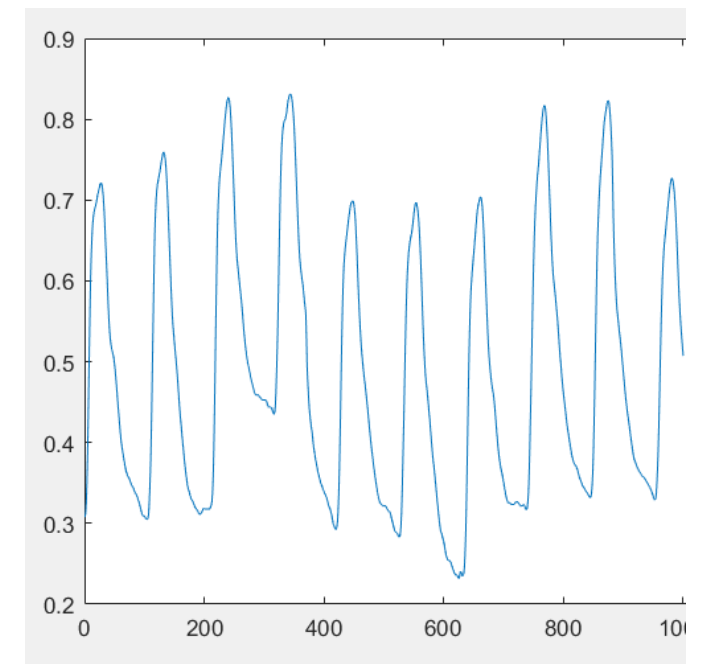
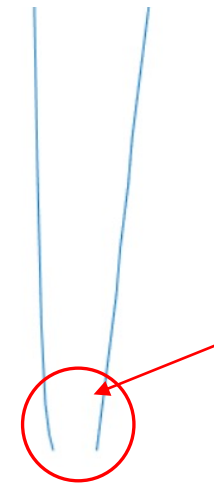
PRÉ-PROCESSAMENTO DE SINAL E EXTRAÇÃO DE FEATURES

Amostras de Sinal PPG



Pré-processamento do sinal

- Todos os sinais foram normalizados entre -1 e 1 (sinal ótico não calibrado, por isso as unidades de intensidade não estão calibradas e podem depender de causas múltiplas e nada relacionadas com o que está a acontecer fisiologicamente);
- Foram retirados valores de sinal que não estavam definidos (remoção de valores ausentes é uma prática comum no pré-processamento de dados para garantir que o conjunto de dados seja consistente e que não haja interrupções ou erros durante a análise);
- Cada sinal foi dividido em segmentos de 8 segundos (suficiente para capturar informações fisiológicas relevantes que normalmente ocorrem em intervalos de vários segundos);



Extração de features

Para cada um dos segmentos de 8 segundos foram calculadas as seguintes features:

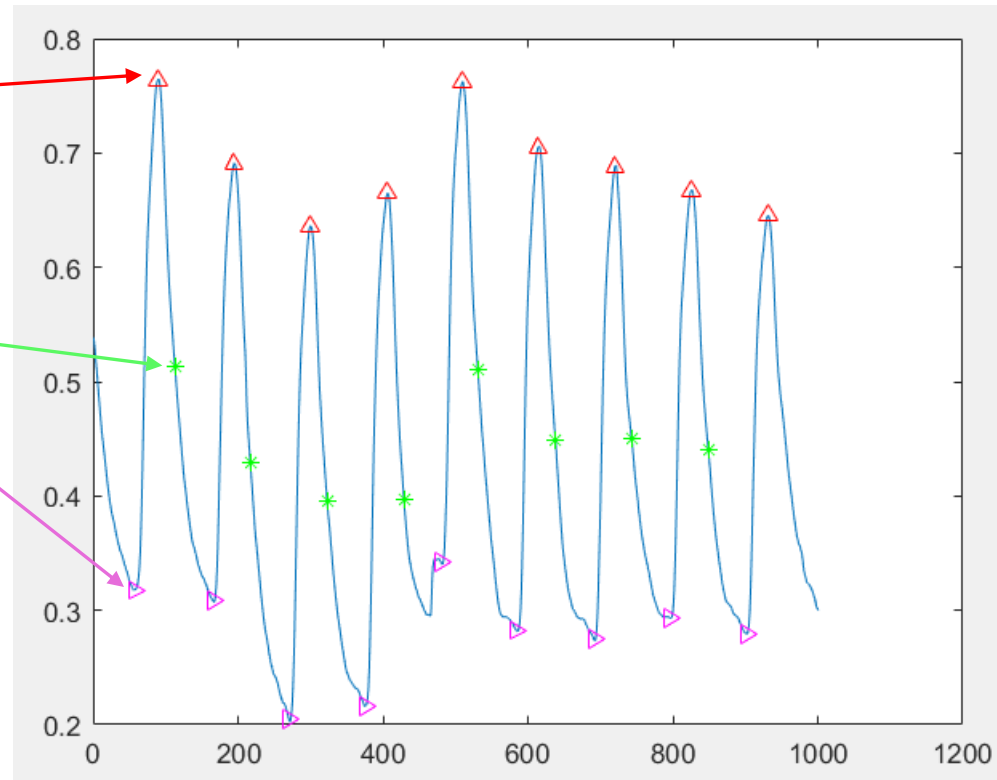
- Estatísticas de domínio temporal:

- Média;
- Mediana;
- Desvio padrão;
- Variância;
- Interquartile Range;
- Range;
- Skewness;
- Kurtosis;
- Raíz média quadrada.

Extração de features

■ Características de forma de sinal:

- Systolic peak;
- Começo de cada batimento;
- Dicrotic Notch.



Normalização e organização das features

Foram criadas duas matrizes de features (uma para treino e uma para teste) onde foram inseridas as features respectivas a cada uma.

Cada matriz possui 9 colunas (uma por feature) e 15000 linhas (uma por segmento).

Cada feature foi normalizada entre -1 e 1.

Seleção de features

Normalmente faz-se uma escolha de quais features usar, de modo a eliminar features redundantes. Esta eliminação permite a redução de dimensionalidade do problema, previne overfitting e muitas vezes melhora a precisão e sucesso dos resultados.

Como neste caso o número de features já era bastante baixo, foram usadas todas as features extraídas.

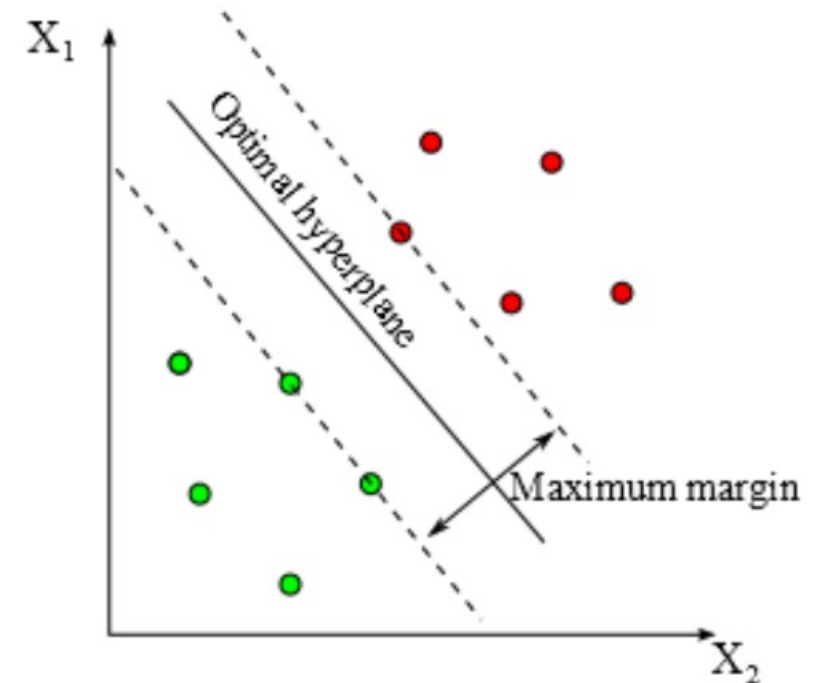
Modelo SVM

TREINO E CLASSIFICAÇÃO

Modelo SVM

Foi usado para este estudo o modelo Support Vector Machine (SVM), que tem como pressuposto, tendo dados de duas classes distintas, encontrar o melhor hyperplano que separa mais eficientemente os dados das duas classes. Este hyperplano é conseguido maximizando a distancia entre os planos que limitam os dados de ambas as classes.

Cada ponto neste espaço de dimensão n representa uma sample de sinal, constituída por um conjunto de n features. Aos pontos que estão no plano limite da classe são chamados Support vectors.



Modelo SVM

A otimização de hiperparâmetros desempenha um papel essencial na determinação do melhor hiperplano que divide as duas classes.

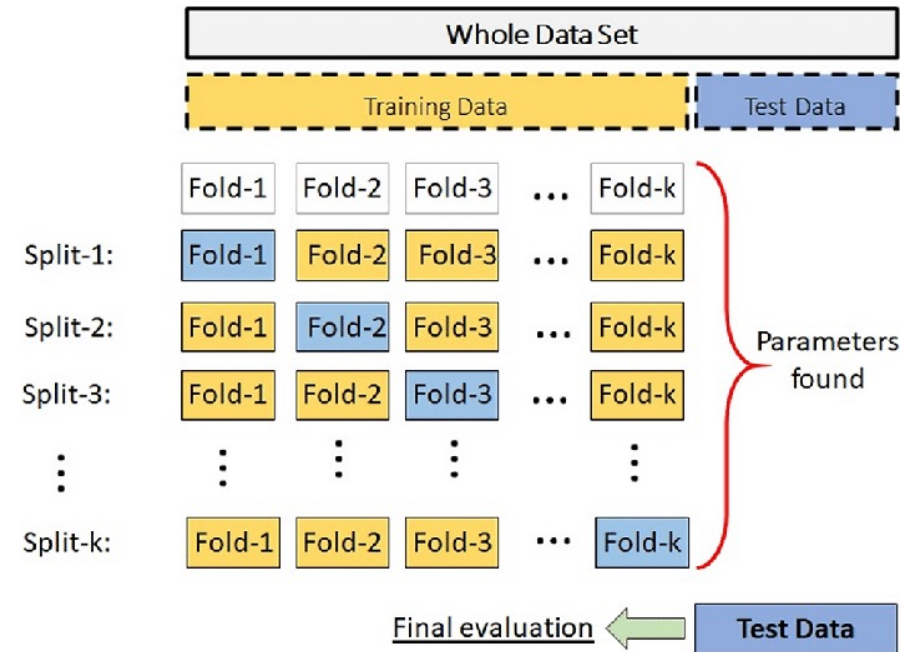
Neste modelo os principais parâmetros são:

- Parametro C: determina o balanço entre tamanho de margem (quanto menor C maior é a margem) e o erro no treino (quanto maior C menor é o erro no treino o que leva também a uma margem menor).
- Kernel Function: Mapeia features não lineares num espaço multidimensional.

Cross Validation

Uma forma de testar a qualidade dos parâmetros escolhidos é usar k fold cross validation.

Este método separa os dados de treino em k folds. Um dos folds é excluído do treino e é usado para testar o modelo obtido. Este processo faz-se iterativamente até que todos os folds tenham servido com fold de teste.



Treino do modelo

Como primeira abordagem usou-se “Radial Basis Function” (RBF), como Kernel Function e para descobrir o melhor conjunto de hyperparametros usou-se Bayesian optimization.

Este processo de otimização foi abandonado uma vez que demorava muito tempo a ser executada e os resultados que produzia, apesar de terem uma exatidão média de 80% quando usando cross-validation, a exatidão quando avaliando o modelo com os dados de teste era de apenas 40%, o que aponta para um caso de overfitting.

Treino do modelo

Como alternativa ao método anterior experimentou-se, iterativamente, diversos valores de C e kernel Scale (parâmetro de escala para os dados de entrada), calculando para cada um a exatidão utilizando k-fold cross validation e usando os dados de teste.

Melhores resultados usando Linear Kernel

Cross Validated model		
C	Kernel Scale	Exatidão
0,001	0,001	77,40%
0,001	0,01	77,26%
0,1	0,01	77,36%
0,1	0,1	77,15%
1	0,1	77,26%
10	0,1	77,24%
10	1	77,14%

Usando dados de Teste		
C	Kernel Scale	Exatidão
0,001	0,001	59,71%
0,001	0,01	80,16%
0,1	0,01	48,62%

Melhores resultados usando RBF

Cross Validated model		
C	Kernel Scale	Exatidão
0,01	0,1	79,09%
0,1	0,1	81,88%
0,1	1	78,27%
1	0,1	84,53%
1	1	80,53%
10	0,1	82,74
10	1	81,81%

Usando dados de Teste		
C	Kernel Scale	Exatidão
0,1	0,1	77,15%
1	0,1	77,26%
10	0,1	77,24%

Treino do modelo

Ambas as kernel functions apresentam resultados semelhantes, no entanto, dados cardíacos com os em questão, não são normalmente lineares. RBF é, já comprovado por estudos anteriores, uma boa função para tratar este tipo de dados.

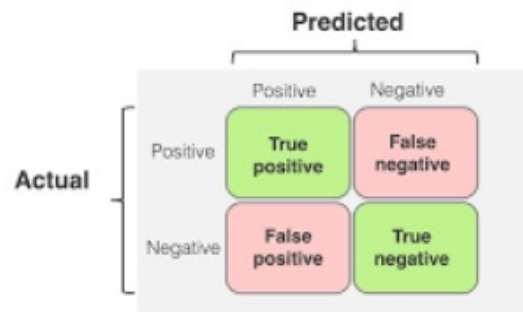
Procurou-se novamente hyperparâmetros que maximizem a exatidão do modelo, partindo dos resultados obtidos anteriormente, usando desta vez apenas RBF.

A melhor combinação de hyperparâmetros encontrado foi:

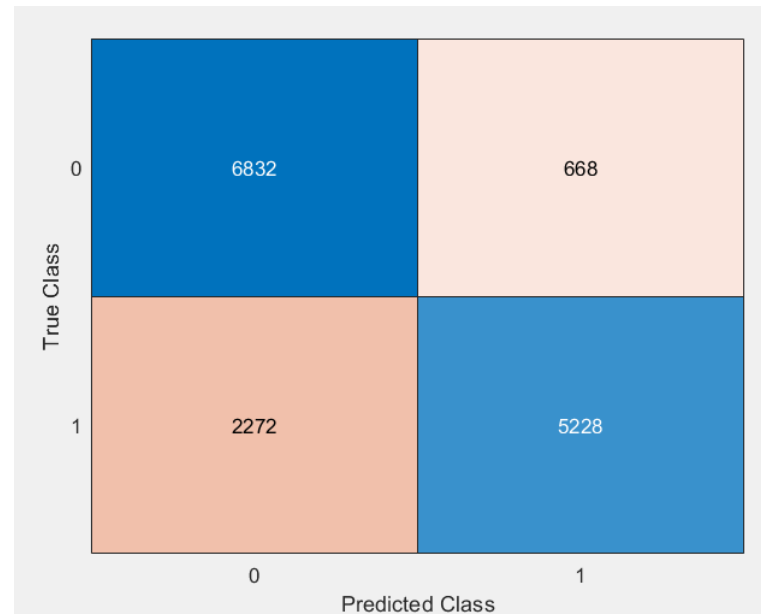
C	Kernel Scale	CrossValidation accuracy	Test accuracy
2,31	1,01	81,12%	80,40%

Confusion Matrix

Uma confusion matrix é normalmente usada para avaliar a performance de um modelo de machine learning.



- $Precis\tilde{a}o = \frac{TP}{TP+FP} = 75\%$
- $Recall = \frac{TP}{TP+FN} = 91\%$
- $F1\ Score = \frac{2*Recall*Precis\tilde{a}o}{Precis\tilde{a}o+Recall} = 82\%$



Conclusões

Podemos verificar que o sucesso das classificações do modelo depende fortemente do conjunto de hiperparâmetros escolhidos e features usadas para treinar o mesmo (ao adicionar as features relativas à distância entre picos a exatidão de classificação subiu cerca de 10%). Foi testado o caso em que treinamos o modelo reduzindo o número de features, mas os resultados foram sempre piores que no caso em que se utilizou a totalidade das features.

Com mais tempo teria sido possível:

- Extrair features mais relevantes e decisivas para a distinção das duas classes;
- Usar métodos mais elaborados de otimização de features (filtragem de sinal indesejado);
- Explorar outras estratégias relevantes na melhoria da qualidade das features;
- Aplicar outros modelos de Machine learning (como Random Forest ou redes neurais) e comparar os resultados obtidos.

Bibliografia

<https://zenodo.org/records/6973963>

<https://www.sciencedirect.com/science/article/abs/pii/S1386505617303726>

https://www.politesi.polimi.it/retrieve/a81cb05c-a20d-616b-e053-1605fe0a889a/Politecnico_di_Milano_Thesis_draft%20%284%29.pdf