Universidade Estadual de Campinas

Instituto de Matemática, Estatística e Computação Científica

Métodos de Otimização Não-Linear e Support Vector Machine

Gabriel Passos RA: 172351
Giovanna Castello de Andrade RA: 168633
João Luiz Santos Gomes RA: 199657
Vinícius Oliveira Martins RA: 206853

27 de Novembro de 2019

1 Introdução

Este trabalho consiste no estudo e implementação dos métodos de **Penalização** e do **Lagrangiano Aumentado**, aprendidos no curso de Programação Não-linear (MS629). Em seguida, apresentamos um modelo de classificação de dados categóricos muito conhecido na área de *machine learning* chamado ***Support Vector Machine(SVM)***. Esta abordagem para o tema foi proposta como exercício em [1, Seção 1.7.2], material que foi utilizado como base para o desenvolvimento deste projeto.

Como motivação, vamos considerar o problema de diagnóstico de cancêr de mama. O objetivo é que sejamos capazes de identificar a presença ou não de câncer malignos em novos pacientes se baseando em dados médicos de pacientes cujos diagnosticos já são conhecidos.

2 Aprendizado de Maquina e Support Vector Machine

De modo geral, vamos supor que temos um conjunto de dados que podem ser classificados de duas formas: possuem ou não possuem certa característica. No contexto de *machine learning*, a área destinada a solucionar este tipo de problema é chamada de reconhecimento ou classificação de padrões.

Vamos supor que temos um conjunto de m dados de treinamento dados por $x_i \in \mathbb{R}^n$, $i = 1, \dots m$. Em termos de classificação, isso significaria que estamos olhando para n variáveis, como por exemplo, o resultado de n exames médicos diferentes, e a partir delas inferindo uma saída $y \in \{1, -1\}$, indicando por exemplo a presença ou ausência de câncer maligno. Dessa forma, a n-úpla $x_i = (x_{i1}, \dots x_{in})$ corresponde aos resultados do i-ésimo paciente. Quando $n = 2$, podemos reprentar o problema graficamente como m pontos no plano, como apresentado na Figura 1. Os pontos vermelhos representam os pacientes cuja saída é $y = 1$, ou seja, em que há a presença de câncer maligno, e os pontos em verde os pacientes cuja saída é $y = -1$, ou seja, em que há câncer benigno.

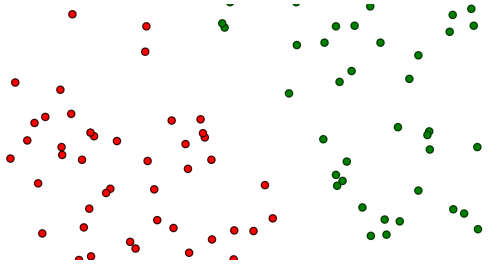


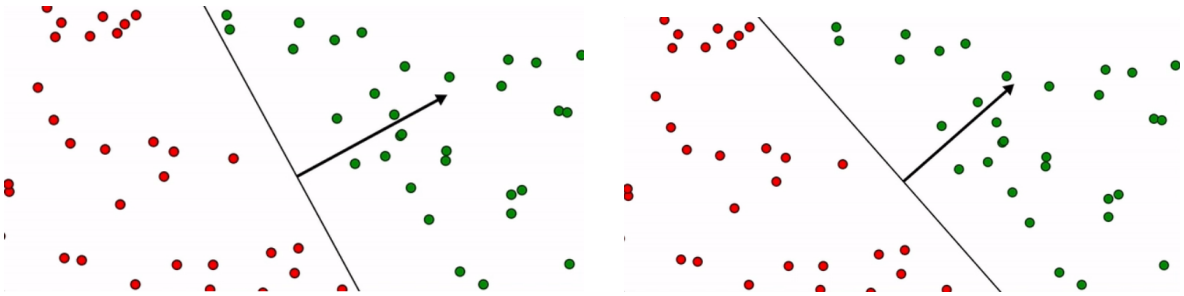
Figure 1: Representação gráfica do conjunto de dados.

O desenvolvimento da técnica de Support Vector Machine parte da ideia de buscar um hiperplano

$$w^t x + b = 0$$

capaz de separar os dados. Uma vez que este plano tenha sido construído, novos dados podem ser classificados a partir de sua localização com relação a ele. Notamos pela Figura

2, no entanto, que mesmo quando este hiperplano existe, ele não é necessariamente único e, portanto, deve haver algum critério para selecionar aquele que melhor separa os dados.



(a) Hiperplano 1. (b) Hiperplano 2.

Figure 2: Duas possibilidades de hiperplanos separando um mesmo conujnto de dados.

Como primeiro critério, vamos supor que existe uma separação mínima entre os pontos positivos e negativos. Formalmente podemos obter isso exigindo que

$$w^t x_i + b \geq +1, y_i = +1$$
$$w^t x_i + b \leq -1, y_i = -1,$$

ou como,

$$y_i(w^t x_i + b) \geq 1.$$

Vamos considerar como hiperplano ideal aquele cuja distância entre os elementos positivos e negativos é a maior possível, já que a boa definição e distinção dos grupos permitirá uma melhor classificação de novos pontos. Para isso, definimos o conceito de **margem de separação** como a distância entre os dois hiperplanos, ilustrada na Figura 3 como *margin*. Vamos demonstrar que esta distância é dada por $\frac{2}{||w||}$.

Sejam dois hiperplanos paralelos $\mathcal{H}_1 : w^t x + b_1 = 0$ e $\mathcal{H}_2 : w^t x + b_2 = 0$ com $x_1 \in \mathcal{H}_1$, $x_2 \in \mathcal{H}_2$. Seja R a reta que passa por x_1 e é perpendicular $\mathcal{H}_1$. Assim a equação de R é $x_1 + w\gamma$, com $\gamma \in \mathbb{R}$. A intersecção de R com $\mathcal{H}_2$ será dada por

$$\begin{aligned} w^t(x_1 + w\gamma) = b_2 &\Leftrightarrow w^t x_1 + (w^t w)\gamma = b_2 \\ &\Leftrightarrow \gamma = \frac{b_2 - w^t x_1}{w^t w} \\ &\Leftrightarrow \gamma = \frac{b_2 - b_1}{||w||^2} \end{aligned}$$

Sendo assim, o ponto de intersecção será $x_2 = x_1 + w\frac{(b_2 - b_1)}{||w||^2}$ e a distância entre os dois hiperplanos será

$$||x_2 - x_1|| = ||w||\left(\frac{|b_2 - b_1|}{||w||^2}\right) = \frac{|b_2 - b_1|}{||w||}.$$

Portanto, tomando $b_1 = b + 1$ e $b_2 = b - 1$ obtemos a expressão desejada.

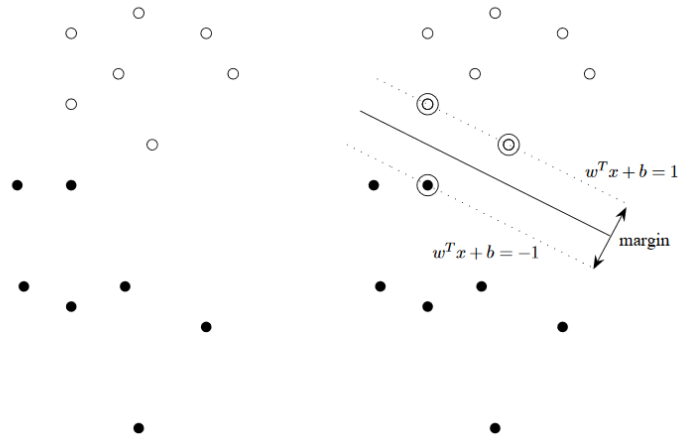


Figure 3: Margem de separação entre os hiperplanos no caso em que os dados são separáveis.

Para que possamos classificar corretamente os grupos de dados, queremos maximizar a distância obtida acima. Isso é equivalente ao problema de minimização abaixo, contendo m restrições lineares de desigualdade, com $x \in \mathbb{R}^n, w \in \mathbb{R}^n, b \in \mathbb{R}$.

$$\begin{aligned} &\text{minimizar} \quad f(w,b) = \frac{1}{2}||w|| \\ &\text{sujeita a} \quad y_i(w^t x_i + b) \geq 1, \quad i = 1, \ldots, m. \end{aligned} \tag{1}$$

Esta formulação, no entanto, nem sempre é suficiente, visto que os dados nem sempre são separáveis, como ilustra a Figura 4. Neste caso, o SVM propõe que seja permitido que pontos violem a separação do hiperplano. No entanto, para toda violação, é acrescentado um termo de penalização positivo na função objetivo. Com isso em vista, quando minimizamos o problema em questão, a solução tenderá a ter o mínimo de violações possível. Assim obtemos o problema

$$\begin{aligned} &\text{minimizar} \quad f(w,b) = \frac{1}{2}||w|| + C\sum_{i=1}^{m} \overline{\xi_i}^2 \\ &\text{sujeita a} \quad y_i(w^t x_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m, \overline{\xi_i} \geq 0. \end{aligned} \tag{2}$$

Para evitar trabalhar com desigualdades, chamamos $\xi_i^2 = \overline{\xi_i}$ e criamos variáveis de folga t_i, de modo que o problema possa ser escrito como

$$\begin{aligned} &\text{minimizar} \quad f(w,b) = \frac{1}{2}||w|| + C\sum_{i=1}^{m} \xi_i^2 \\ &\text{sujeita a} \quad y_i(w^t x_i + b) - 1 + \xi_i^2 - t_i^2 = 0, \quad i = 1, \ldots m. \end{aligned} \tag{3}$$

É interessante notar que as variáveis ξ_i indicam o quanto o ponto x_i violou a posição esperada em relação ao hiperplano obtido, enquanto as variáveis t_i indicam se cada restrição foi ou não satisfeita na igualdade.

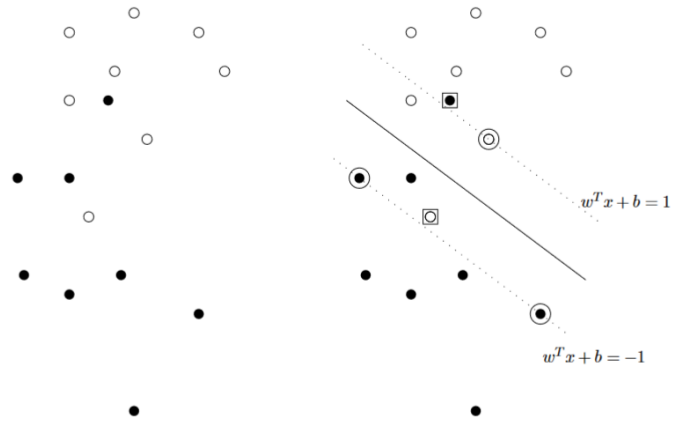


Figure 4: Caso em que os dados não são separáveis.

2.1 Implementação e Solução via Método da Penalização

Utilizando a função de penalização quadrática definimos a seguinte função

$$\phi(\rho_k, w, b) = \frac{1}{2}||w|| + C\sum_{i=1}^{m}\xi_i^2 + \rho_k\sum_{j=1}^{m}[y_j(w^t x_j + b) + -1 + \xi_j^2 - t_i^2]^2.$$

Tomando agora uma sequencia $(\rho_k)_{k\in\mathbb{N}}$ tal que $\rho_{k+1} > \rho_k$, obtemos os subproblemas:

$$beginaligned\text{minimizar}\phi(\rho_k, w, b, \xi, t)\text{sujeita a}w \in \mathbb{R}^n, b \in \mathbb{R}, \xi_i \in \mathbb{R}^m, t_i \in \mathbb{R}^n$$

(4)

cujas soluções fornecem uma sequência (w_k, b_k) de soluções que convergem para a solução (w_k^*, b_k^*) do problema original (teorema 12.1.1 de [4]). A seguir, vamos calcular os gradientes da função , que serão utilizados na implementação do algoritimo

$$\nabla_w\phi = w + 2\rho\sum_i h_i\nabla_w h_i$$

$$\nabla_\xi\phi = 2C\xi + 2\rho\sum_i h_i\nabla_\xi h_i$$

$$\nabla_t\phi = 2\rho\sum_i h_i\nabla_t h_i$$

$$\nabla_b\phi = 2\rho\sum_i h_i\nabla_b h_i$$

Algoritmo 1. Método da Penalização

Seja $z_0 = (w_{0,0}\,, t_0, b_0) \in \mathbb{R}^{n+m+m+1}$ um vetor inicial, $\rho_0 > 0, k = 0$.

Enquanto $k < it_{max}$ ou *critério de convergência*, faça:

1. Resolver o subproblema 2.1 para obter z_{k+1} em função de z_k e ρ_k.
2. Atualização do ρ, de forma que $\rho_{k+1} = \alpha\rho k$, com $\alpha > 1$.

Apesar do Método da Penalização possuir uma teoria bem estruturada este método possui algumas dificuldades numéricas. O subproblema pode se tornar extremamente difícil de ser resolvido, isto ocorre pois quando ρ_k tende ao infinito os autovalores da Hessiana $\nabla^2\phi(\rho_k, w, b, \xi, t)$ se aproximam de infinito o que torna essa matriz mal condicionada.

Assim, o método do lagrangiano aumentado se mostra mais eficiente para problemas de grande porte.

2.2 Implementação e Solução via Método do Lagrangiano Aumentado

Para implementação do método do Lagrangiano Aumentado, vamos considerar o seguinte problema de minimização

$$\begin{array}{ll} \text{minimizar} & L_A(\rho_k, w, b, \xi, t) \\ \text{sujeita a} & w \in \mathbb{R}^n, b \in \mathbb{R}, \xi_i \in \mathbb{R}^m, t_i \in \mathbb{R}^n \end{array} \tag{5}$$

em que

$$L_A(\rho_k, \lambda_k, w, b, \xi, t) = \frac{1}{2}||w|| + C\sum_{i=1}^{m}\xi_i^2 + \sum \lambda_i^k h_i(w, \xi, t, b) + \rho_k \sum_{j=1}^{m} h_i(w, \xi, t, b)^2$$

e $y_j(w^t x_j + b) + -1 + \xi_j^2 - t_i^2$.

Computando o gradiente da função L_A e das restrições h_i, temos:

$$\begin{aligned} \nabla_w L_A &= w + \sum \lambda_i^k \nabla_w h_i + 2\rho \sum h_i \nabla_w h_i \\ \nabla_\xi L_A &= 2C\xi + \sum \lambda_i^k \nabla_\xi h_i + 2\rho \sum h_i \nabla_\xi h_i \\ \nabla_t L_A &= \sum \lambda_i^k \nabla_t h_i + 2\rho \sum h_i \nabla_t h_i \\ \nabla_b L_A &= \sum \lambda_i^k \nabla_b h_i + 2\rho \sum h_i \nabla_b h_i \end{aligned}$$

$$\nabla_w h_i = y_i x_i, \quad \nabla_x i h_i = 2C\xi\hat{e}_i, \quad \nabla_t h_i = -2t\hat{e}_i, \quad \nabla_b h_i = y_i.$$

em que $\hat{e}_i$ é o i-ésimo vetor canônico.

Dessa forma, podemos construir os vetores gradiente $\nabla h_i = (\nabla_w h_i, \nabla_\xi h_i, \nabla_t h_i, \nabla_b h_i)^t, \nabla L_A = (\nabla_w L_A, \nabla_\xi L_A, \nabla_t LA, \nabla_b L_A)^t$.

Para a implementação do método, trabalharemos com um vetor de variáveis $z = (w, , t, b) \in \mathbb{R}^{n+m+m+1}$.

Algoritmo 2. Método do Lagrangeano Aumentado

Seja $z_0 = (w_{0,0}\,, t_0, b_0) \in \mathbb{R}^{n+m+m+1}$ um vetor inicial, $\lambda_0, \rho_0 > 0, k = 0$.

Enquanto $k < it_{max}$ ou *critério de convergência*, faça:

1. Resolver o subproblema 5 para obter z_{k+1} em função de z_k, ρ_k e λ_k.
2. Atualização do λ, de forma que $\lambda_{k+1} = \lambda_k + \rho_k h(x_k)$
3. Atualização do ρ, de forma que $\rho_{k+1} = \alpha\rho k$, com $\alpha > 1$.

Fim do enquanto.

3 Resultados

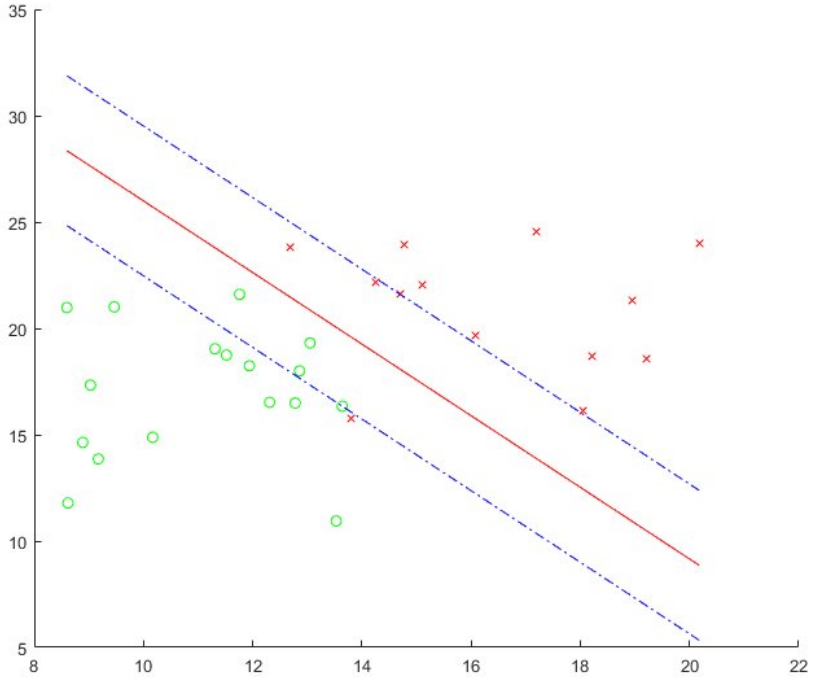
Como proposto em [1], o algoritimo de Support Vector Machine foi aplicada para solução do problema de diagnostico de câncer maligno ou benigno em pacientes com câncer de mama. Para isso, foi utilizada a base de dados disponível em `http://math.gmu.edu/~igriva/book/ch1data.html`.

Para resolução dos subproblemas de minimização irrestrita, foi utilizada a função `fminunc` do `MATLAB,` destinada à obtenção de solução de problemas de otimização irrestrita com várias variáveis.

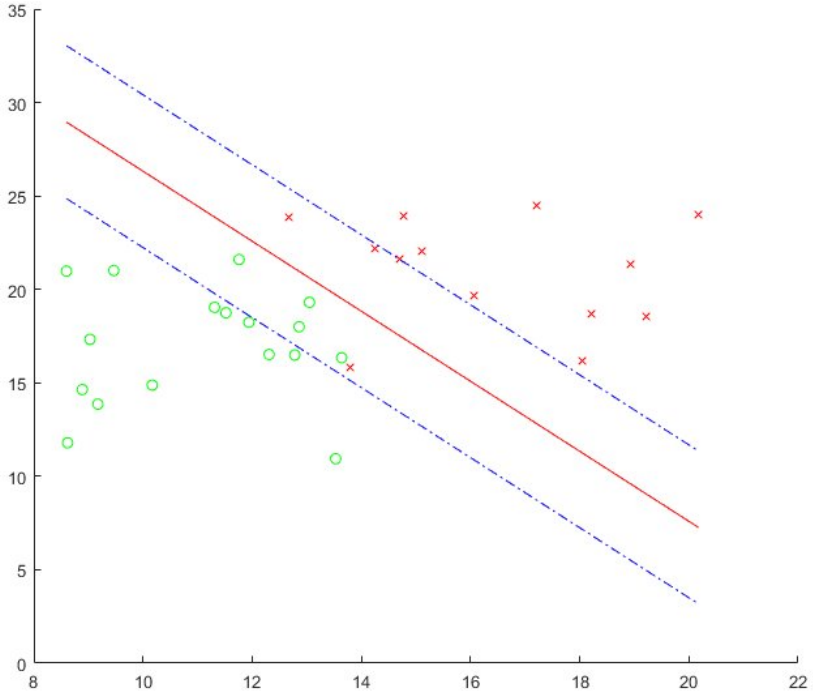
Para que fosse possível visualizar graficamente os resultados da forma como são apresentados nas Figuras 3 e 4, foram consideradas apenas as duas primeiras variáveis, ou seja, a entrada dos dois primeiros exames de cada uma das m pacientes. As Figuras 6a e 6b apresentam os resultados obtidos quando o método foi executado apenas para os 30 primeiros pacientes. Notamos que os resultados obtidos para os dois métodos são bastante similares, ainda que diferentes. As retas tracejadas no gráfico correspondem aos hiperplanos $w^t x + b = 1$ e $w^t x + b = -1$, enquanto as retas vermelhas correspondem ao hiperplano $w^t x + b = 0$. Os pontos vermelhos correspondem aos pacientes com câncer maligno e os pontos verdes aos pacientes com câncer benigno.

Notamos que de acordo com a formulação do nosso problema, queremos que os dois grupos de pontos estejam o mais separados o possível. Além disso, queremos o menor número de pontos possível violem a restrição. Graficamente é possível visualizar que ocorrem mais violações na solução com o método da penalização.

No item 7.2 de [1], o autor propõe que os dados dos 500 primeiros pacientes sejam utilizados para treinar o modelo, ou seja, gerar um hiperplano. Uma vez que isso tenha sido feito, o diagnóstico dos 69 pacientes restantes podem ser definido a partir deste hiperplano e o resultado pode ser comparado com o diagnóstico real. Com isso em mente, o procedimento descrito acima foi repetido para os 500 pacientes, resultando nos gráficos das Figuras 6a e 6b.

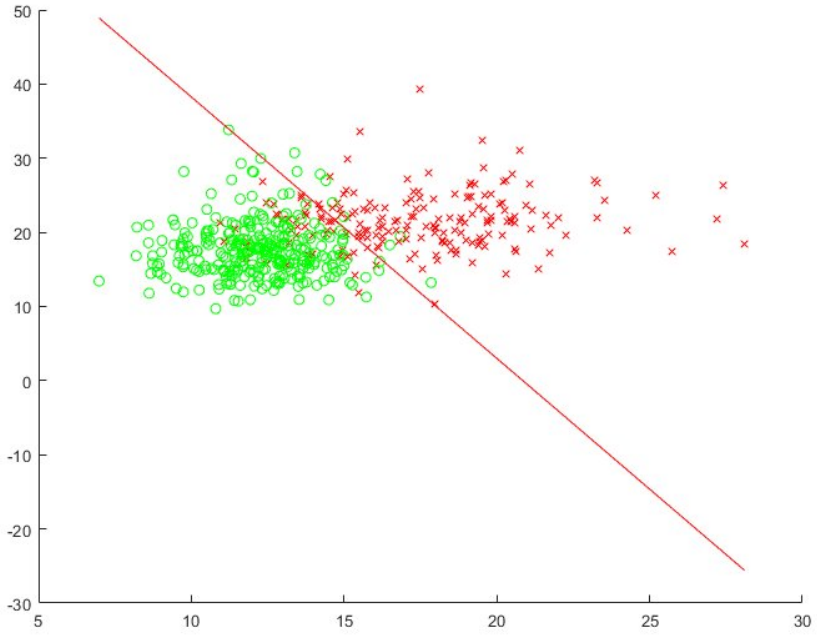


(a) Método do Lagrangiano Aumentado

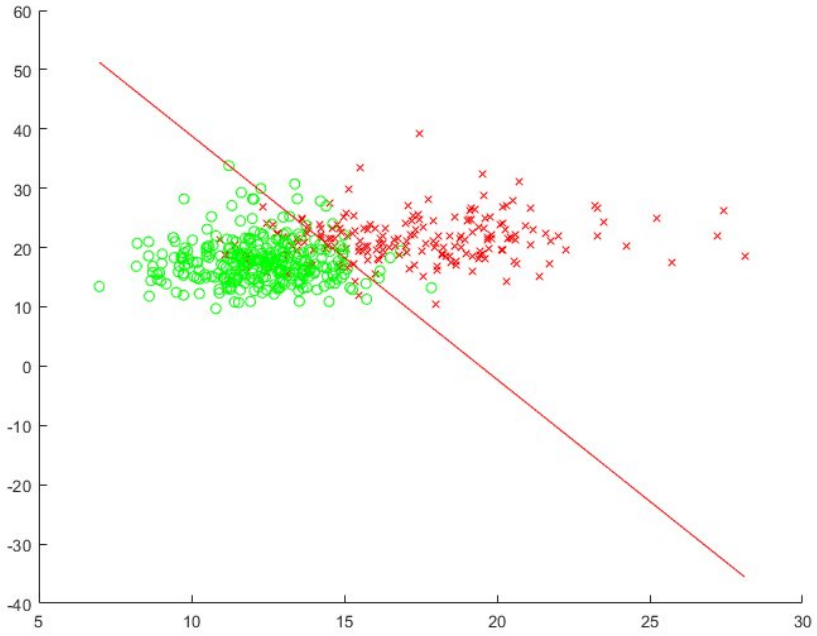


(b) Método da Penalização

Figure 5: Resultados obtidos para dados dos 30 primeiros pacientes.



(a) Método do Lagrangiano Aumentado



(b) Método da Penalização

Figure 6: Resultados obtidos para dados dos 500 primeiros pacientes.

Avaliando os dados dos pacientes restantes, ou seja, dos vetores x_i, $i = 500, \ldots, 569$, na equação $w^t x + b$ do hiperplano, obtivemos uma taxa de acertos de $97,1\%$ para o método do Lagrangiano Aumentado e $91,30\%$ para o método da Penalização. É importante ressaltar que a decisão entre câncer maligno e benigno a partir da equação do hiperplano foi feita da seguinte forma: caso $w^t x_i + b > 0$, então estamos acima do hiperplano e o diagnóstico é de câncer maligno. Caso contrário, o câncer é benigno.

Dessa forma, concluímos que a solução do problema via método do Lagrangiano Aumentado forneceu resultados mais satisfatórios que o método da penalização. De toda forma, apesar de ambos os métodos terem apresentado uma taxa de acerto alta, é importante ressaltar que para este relatório os resultados foram gerados utilizando apenas as duas primeiras variáveis em questão, ou seja, dados de 2 exames apenas. Caso fossem utilizados os exames restantes, é esperado que a taxa de acertos aumente.

References

[1] Griva, I.; Nash, S. G. & Sofer, A. (2008), *Linear and Nonlinear Optimization* (2. ed.), SIAM .

[2] Friedlander, A. Elementos de Programação Não-linear. Disponível em `https://www.ime.unicamp.br/~friedlan/`.

[3] Haykin, S. Neural Networks and Learning Machines (1999). Third Edition. Editora Pearson, Ontario, Canada.

[4] Fletcher, R. Practical methods of optimization (2000). Second Edition. Editora Wiley.