Otimização contínua – Trabalho

Ano letivo: 2024/25 25 de novembro de 2024

Data de entrega - 23h59, 13 de dezembro de 2024

Regras e indicações para o trabalho

1. O trabalho pode ser feito só ou em pares. Se forem realizar o trabalho em pares quero que me indiquem o vosso grupo até ao final desta semana (29/11).

2. Podem usar qualquer linguagem não demasiado exótica, ou qualquer programa que permita código simples. Python será possivelmente o mais popular entre vós, mas se quiserem usar R, matlab, C, julia,... estão à vontade. Se for algo muito exótico consultem-me primeiro.

3. Podem falar entre vocês, trocar ideias e pedir ajuda uns aos outros (o que é diferente de copiar o trabalho). Peço-vos no entanto que no vosso relatório incluam uma lista das pessoas que vos ajudaram.

4. Também quero que o relatório inclua o número de horas que gastaram no trabalho. Se o trabalho for feito por um par quero o número de horas separado por cada pessoa e saber quem fez o quê exatamente.

5. Sobre LLM's. Muitos de vocês usam ChatGPT, Claude, e outros que tais no dia a dia. Não me oponho a que as usem neste trabalho se quiserem/precisarem mas há regras:

   (a) Todo e qualquer uso de LLM's deve ser discriminado no relatório dizendo o que foi feito com recurso a LLM's. Por exemplo: "melhorei a qualidade do texto final com ChatGPT", ou "usei o Claude para gerar um primeiro esboço do algoritmo de descida do gradiente estocástico", ou "usei o Gemini para traduzir o enunciado",...

   (b) A responsabilidade do que está no relatório é vossa. A desculpa: *foi o ChatGPT que disse isso*, não convence ninguém. Exijo que percebam totalmente todas as linhas que estiverem no trabalho que entregarem, exatamente como se as tivessem escrito sem auxílio de LLM's. E vou pedir-vos para explicar com certeza.

   (c) Esteticamente a escrita gerada por LLM's tende a ser pedante e demasiado floreada. Tenham atenção a isso.

   (d) Usem espírito crítico. Se usarem LLM's de forma cega e generalizada, nunca aprenderão as bases necessárias a perceber se o que as LLM's vos estão a dizer faz ou não sentido e tem qualidade. O resultado final sofrerá. E nunca se esqueçam, se não sabem fazer nada que uma LLM não faça então as empresas vão contratar a LLM e não a vocês!

6. Podem pedir-me ajuda a mim, não serão penalizados por isso, mas não deixem para o último dia porque posso não ter tempo de vos ajudar a todos.

7. Quero que me entreguem o código e, à parte, um relatório de texto em formato pdf, onde mostram os resultados que obtiveram num formato fácil de ler e não demasiado extenso. O relatório terá no máximo quatro páginas (e pode ter menos sem problema). O relatório será o principal elemento de avaliação.

8. Não preciso que o código que façam seja muito elaborado. Prefiro algo simples, que compreendam bem, e sem grandes sofisticações. Os algoritmos que vos peço para implementar não ocupam mais de meia dúzia de linhas cada. Estou mais interessado em como é que os usam de que em como os implementam. Esta não é uma disciplina de programação. Isso não impede que queira que comentem o código minimamente para perceber o que estão a fazer.

9. Tenham em atenção que há muitas versões da regressão logística. Se consultarem fontes aleatórias na internet, ou LLM's, ou outras fontes quaisquer, a probabilidade de vos estarem a dizer como fazer algo semelhante mas não exatamente igual ao que eu quero é grande. Tenham cuidado de resolver o que eu peço e não o que a internet julga que eu quero.

10. Uma ressalva: este trabalho não é realmente fiel àquilo que são as melhores práticas. Tipicamente não iriam usar um método de primeira ordem para resolver uma regressão logística e fariam um tratamento um pouco mais sofisticado dos dados antes de otimizar. Mas serve de ilustração ao que está por trás dos métodos pré-implementados a que têm acesso.

11. Planeio realizar as defesas na semana seguinte à entrega (16,17 e 18 de dezembro). exatamente como dividimos e calendarizamos as defesas será uma questão a ver mais tarde, quando souber quantos trabalhos vão existir para serem defendidos.

12. Alguma coisa que não esteja esclarecida nestas indicações, perguntem-me diretamente.

Apresentação do dataset - WBCD

No ficheiro *wdbc.csv* disponível no material de apoio encontram dados de 569 diagnósticos de cancro da mama ("Breast Cancer Wisconsin (Diagnostic)" (WBCD) dataset, mantido pelo UCI - Machine Learning Repository). Cada linha representa um caso, sendo as primeiras 30 colunas parâmetros medidos em amostras e a coluna final contendo o diagnóstico -1 ou 1 para benigno ou maligno respetivamente. Um plot do dataset pode ser visto na Figura 1. O nosso dataset é assim constituído por 569 pares (x_i, y_i), com $x_i \in \mathbb{R}^{30}$ e $y_i \in \{-1, 1\}$.

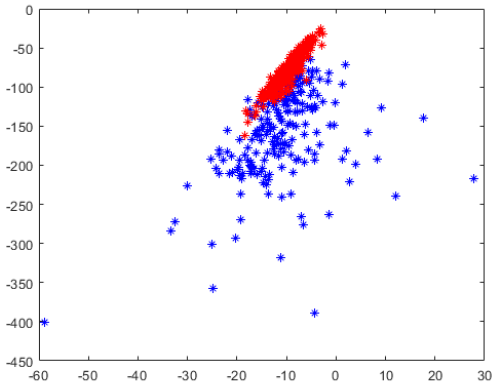

Figura 1: Plot do dataset wdbc projetado em 2D

Vamos usar estes dados para o trabalho. A ideia será usar os nossos algoritmos de otimização para treinar classificadores binários usando a regressão logística. O nosso objetivo é encontrar um $a \in \mathbb{R}^{30}$ e um $b \in \mathbb{R}$, de forma a que dado uma nova amostra $\overline{x}$ se vá classificar com 1 se $a^t\overline{x} > b$ e com -1 se $a^t\overline{x} < b$. Para não ter de tratar o a e o b separadamente, é usual acrescentar um 1 ao início de cada x_i, ficando assim com um vector em $\mathbb{R}^{31}$. O classificador linear anterior passa a ser equivalente a encontra $a \in \mathbb{R}^{31}$ tal que classificamos com 1 se $a^t x > 0$ e -1 caso contrário.

Passo 0 - Preparar os dados

1. Carrega os dados e acrescenta um 1 no inícios dos x_i, transformando-os em vetores em $\mathbb{R}^{31}$.

2. Usando um método à tua escolha, faz a tua versão da Figura 1. Isto é, usando um método à tua escolha, faz um plot 2D dos 569 x_i's, marcando de forma diferente os benignos dos malignos.

3. Escolhe aleatoriamente 450 pontos para serem o teu conjunto de treino, deixando os restantes 119 para teste. (Não vamos ter conjunto de validação). Repete o plot anterior para cada um dos dois conjuntos, para verificar se parecem bem distribuídos.

Passo 1 - Regressão logística

Na versão mais simples da regressão logística o classificador linear encontra-se minimizando a soma

$$f(a) = \frac{1}{450} \sum_{i=1}^{450} \ln\left(1 + e^{-y_i a^T x_i}\right)$$

para $a \in \mathbb{R}^{31}$ (vejam as notas da disciplina para mais detalhes). A derivada desta função é

$$\nabla f(a) = \frac{1}{450} \left(\sum_{i=1}^{450} \frac{-y_i}{1 + e^{y_i a^T x_i}} x_i \right).$$

1. Implementa e utiliza o método da descida do gradiente de passo fixo para resolver o problema de regressão logística para estes dados. Encontra um tamanho do passo e um número de iterações adequado, justificando a tua escolha.

2. Implementa e utiliza agora a descida de gradiente estocástica de passo fixo para resolver o mesmo problema. Compara a convergência com a convergência do método do gradiente.

3. O a encontrado em ambos os métodos deve ser semelhante. Usa o obtido na primeira alínea para classificar os 169 pontos do conjunto de teste. Quantos estão bem classificados? Faz o plot da classificação obtida e compara-a com a classificação correta.

Passo 2 - Regressão logística regularizada

Vamos agora considerar o problema da regressão logística com um regularizador l_1. Queremos minimizar:

$$g(a) = \frac{1}{450} \sum_{i=1}^{450} \ln\left(1 + e^{-y_i a^T x_i}\right) + \alpha \|a\|_1$$

onde α é uma constante positiva fixa.

1. Implementa e utiliza o método da descida do gradiente proximal de passo fixo para resolver o problema de regressão logística para estes dados.

2. Usa o algoritmo para resolver o problema anterior com vários valores de α. O que acontece à solução obtida quando o valor de α aumenta? (compara por exemplo o número de entradas não nulas de a e o número de pontos bem classificados)

3. Escolhe um α que te pareça uma escolha razoável e compara a classificação obtida com a obtida anteriormente.

Passo Opcional - scikit-learn

A função sklearn.linear_model.LogisticRegression da biblioteca scikit-learn é uma implementação da regressão logística. Como esta, existem outras.

1. Usa esta ou outra implementação já existente para resolver a regressão logística e compara os resultados com os que obtiveste anteriormente.

Esta alínea não é obrigatória e servirá apenas para compensar alguma falha nas alíneas anteriores.