IA525/EG425 - Otimização Inteira e Combinatória

Atividade 05: Classificadores Lineares | Prof. Matheus Souza

Instruções Gerais

- Esta atividade deve ser resolvida **individualmente**.
- Os itens teóricos devem resolvidos de forma organizada, clara e formal.
- A solução encontrada deve ser submetida, em um único arquivo PDF, no `moodle`. Certifique-se de que todas as resoluções digitalizadas estão legíveis antes de submetê-las.
- Entregas após o prazo estabelecido no `moodle` serão desconsideradas.
- É permitida a consulta a livros e outros materiais, mas a atividade apenas pode ser discutida com a equipe de ensino.
- Os algoritmos desenvolvidos nos itens práticos devem ser organizados e comentados. Todos os códigos utilizados devem ser submetidos como anexos no `moodle`.
- Qualquer tentativa de fraude, se detectada, implicará na reprovação (com nota final 0.0) de todos os envolvidos, além das penalidades disciplinares previstas no Regimento Geral da Unicamp (Arts. 226 – 237).

Apresentação

O problema de classificação é mais uma clássica aplicação da otimização que tem recebido intensa atenção atualmente. Em um problema de classificação binária, temos acesso a um conjunto de *atributos* $t^{(i)} \in R^n$, $i = 1, \cdots, N$, que estão associados a duas classes: verdadeiro ($y^{(i)} = 1$) e falso ($y^{(i)} = -1$). O classificador é um modelo $\phi : \mathbb{R}^n \to \{-1, 1\}$ tal que $\hat{y}^{(i)} = \phi(t^{(i)})$ seja uma estimativa para $y^{(i)}$.

O problema de projeto de um classificador linear via quadrados mínimos envolve a definição de $a \in \mathbb{R}^n$ e $b \in \mathbb{R}$ de forma que o resíduo

$$f(x) = \sum_{i=1}^{N} \Big(y^{(i)} - a^\top t^{(i)} - b \Big)^2$$

seja minimizado. Outra forma de projetar um classificador linear consiste em considerar a típica formulação com *soft margin* de máquinas de vetores de suporte (SVMs), que busca $a \in \mathbb{R}^n$ e $b \in \mathbb{R}$ que minimizam uma função conhecida por *hinge loss*, que penaliza os pontos mal-classificados, sendo dada por

$$f(x) = \sum_{i=1}^{N} \max \Big\{ 0, 1 - y^{(i)}(a^\top t^{(i)} + b) \Big\}.$$

Este segundo problema pode ser resolvido com técnicas de programação linear. Nos dois casos, determinados a, b, temos que o classificador é dado por

$$\phi(t) = \operatorname{sgn}\Big(a^\top t + b \Big) \in \{-1, 1\}.$$

Nesta tarefa, desejamos desenvolver e testar classificadores de três rótulos de forma a classificar flor de Iris em três espécies: *iris setosa*, *iris versicolor* e *iris virginica*, que serão denominadas de classes 1, 2 e 3, respectivamente. O classificador utiliza $n = 4$ atributos, a saber: comprimento da sépala, largura da sépala, comprimento da pétala e largura da pétala. Os dados fornecidos são clássicos, e envolvem 150 exemplos publicados em 1936 pelo estatístico Ronald Fisher. No conjunto de dados fornecidos, escolha 100 exemplos para treinamento e os demais 50 exemplos para teste. Explique como esta escolha foi feita.

Questões

- **Questão 1:** Considerando cada uma das duas técnicas descritas acima, encontre três classificadores binários, cada um classificando uma espécie contra as outras duas (logo, você fará seis classificadores no total, três por QM e outros três por SVM). Forneça a taxa de erro de cada um destes classificadores, tanto no conjunto de treinamento quanto no conjunto de teste.
- **Questão 2:** Combine os classificadores desenvolvidos acima para obter dois classificadores de 3 classes, um desenvolvido por quadrados mínimos e o segundo por otimizaçao linear, e forneça a matriz de confusão para os conjuntos de treinamento e de teste.
- **Questão 3:** Compare os resultados obtidos nos itens anteriores. Qual classificador apresentou melhor desempenho, em geral?