Primeira Lista de Exercícios da Disciplina "Aprendizagem de Máquina"

André da Motta Salles Barreto
amsb@lncc.br

Data de entrega: 03/11/2009

Foram criados cinco conjuntos de dados, dos quais dois se referem a problemas de regressão e três são problemas de classificação. Os dados estão divididos em um conjunto de treinamento, em que constam os valores das variáveis de entrada ($\mathbf{x}$) e das variáveis de saída (y), e um conjunto de teste, com apenas os valores de $\mathbf{x}$. O objetivo dos exercícios é prever o valor da variável y para os dados de entrada $\mathbf{x}$ dos conjuntos de teste.

Os arquivos com os dados estão disponíveis na página da lista de discussões da disciplina. Os nomes dos arquivos são formados da seguinte maneira:

Tipo	+	**N^o**	+	**Trein. / Teste**	+	**Entrada / Saída**
`reg` / `class`		$\{1, 2, 3\}$		`tr` / `ts`		`X` / `Y`

Por exemplo, os dados de saída do conjunto de treinamento do terceiro problema de classificação estão no arquivo `class_3_tr_Y.dat`, em que a extensão "dat" foi usada para indicar "dados". Os dados estão organizados de forma tabular, com um exemplo de treinamento por linha e um atributo por coluna. As colunas estão separadas por espaços em branco.

Para cada conjunto de dados, o aluno deverá escolher um modelo de acordo com as características do problema e implementar um algoritmo de treinamento para configurar os parâmetros do modelo escolhido. A implementação poderá ser feita na linguagem de programação de preferência do aluno. O aluno deverá, então, *para cada um dos cinco problemas*:

1. Apresentar uma justificativa para a escolha do modelo.

2. Nos casos dos problemas de regressão, fornecer a média dos erros quadráticos no conjunto de treinamento, dada por:

$$\frac{1}{m}\sum_{i=1}^{m}(h(\mathbf{x}_i) - y_i)^2,$$

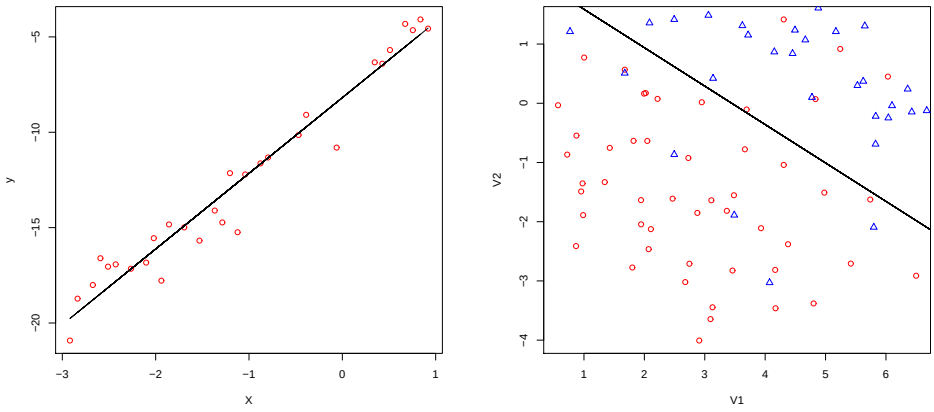
onde $h(\mathbf{x}_i)$ é a resposta dada pelo modelo configurado quando alimentado com a variável $\mathbf{x}_i$ e m é o número de exemplos de treinamento. Para os problemas de regressão, fornecer a taxa de erro de classificação, dada por:

$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\{h(\mathbf{x}_i) \neq y_i\},$$

onde $\mathbf{1}\{\text{Verdadeiro}\} = 1$ e $\mathbf{1}\{\text{Falso}\} = 0$ e $h(\mathbf{x}_i)$ é a classe atribuída à $\mathbf{x}_i$ pelo modelo configurado.

3. Fornecer os valores encontrados para os parâmetros do modelo. Por exemplo, no caso da regressão linear e da regressão logística o aluno deve reportar o valor encontrado para o vetor θ; no caso do classificador Naive Bayes devem ser fornecidos os valores dos parâmetros ϕ_y, $\phi_{i|y=1}$ e $\phi_{i|y=0}$, e assim por diante. Se o modelo escolhido pelo aluno for não-paramétrico, explicar como as variáveis que definem o comportamento do algoritmo de treinamento foram determinadas. Por exemplo, na regressão linear ponderada localmente (LWR), explicar como a vizinhança foi definida.

4. Para os problemas de regressão, gerar um gráfico com os dados de treinamento e a curva computada pelo modelo, nos moldes do que mostra a Figura 1a. Para os problemas de classificação 1 e 2, mostrar um gráfico com os dados de treinamento e a superfície de separação computada pelo modelo, como na Figura 1b.



(a) Problema de regressão (b) Problema de classificação

Figura 1: Ilustração de como devem ser aproximadamente as figuras geradas no exercício

5. Mostrar o código-fonte usado no exercício. O código deve estar comentado e conter apenas o estritamente necessário para o execício em questão.

6. Enviar por e-mail um arquivo com os valores da variável y encontradas para os dados de entrada $\mathbf{x}$ do conjunto de teste (ou seja, para cada $\mathbf{x}_i$ do conjunto de teste, fornecer o y_i correspondente). Os valores de y reportados pelo aluno serão comparados com os valores reais dessa variável, que não foram fornecidos com os exercícios. Os dados devem estar organizados da mesma maneira que os arquivos da lista, com um valor de y em cada linha. Os nomes dos arquivos devem seguir a seguinte lei de formação:

Sobrenome do aluno	+	**Tipo**	+	**N^o**
		`reg` / `class`		$\{1, 2, 3\}$

.

Por exemplo, as minhas respostas para o segundo problema de regressão deveriam estar contidas em um arquivo denominado `barreto_reg_2.dat`.