Lista de Exercícios referente ao Cap. 3 do livro do Bishop

Exercícios Extras (computacionais):

E1) Considere o experimento computacional denominado "Polynomial Curve Fitting", usado diversas vezes no livro texto (veja páginas 4 e 5 do livro, bem como Apêndice A), considerando a ordem do modelo sendo M = 9 e o tamanho da amostra sendo N = 10.
Faça:
(a) Calcule a solução de mínimos quadrados (LS) $\mathbf{w}_{LS}$;
(b) Calcule a solução via regressão ridge (escolha um fator de regularização razoável) $\mathbf{w}_{ridge}$;
(c) Calcule a solução via regressão lasso (escolha um fator de regularização razoável) $\mathbf{w}_{lasso}$;
(d) Monte uma tabela exibindo os 10 coeficientes **w** para as 3 soluções obtidas nos itens acima e comente/compare os resultados;
(e) Plote uma figura contendo o processo gerador em verde (a senoide), e suas estimativas $\mathbf{y}_{LS}$, $\mathbf{y}_{ridge}$ e $\mathbf{y}_{lasso}$ em preto, azul e vermelho, respectivamente.
(f) Repita todos os itens anteriores para N=20 e N=50.

E2) (Dados de câncer de próstata)
Data Source:
- (info) https://web.stanford.edu/~hastie/ElemStatLearn/datasets/prostate.info.txt
- (database) https://web.stanford.edu/~hastie/ElemStatLearn/datasets/prostate.data

"The data for this example come from a study by Stamey et al. (1989). They examined the correlation between the level of prostate-specific antigen (lpsa) and a number of clinical measures in men who were about to receive a radical prostatectomy. The variables are log cancer volume (lcavol), log prostate weight (lweight), age, log of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45)."

Considere a variável (lpsa) como 'target' e as variáveis (lcavol), (lweight), (age), (lbph), (svi), (lcp), (gleason) e (pgg45) como 'entradas'.
Siga o roteiro abaixo:

(a) Padronize os atributos de entrada para que eles tenham média 0 e variância 1;

(b) Divida o dataset em dois conjuntos, treinamento e teste, conforme indicado nos índices da última coluna (T = treinamento, F = teste);

(c) Encontre o modelo linear de regressão ótimo no critério de mínimos quadrados (solução LS);

(d) Implemente modelos lineares regularizados pelos métodos 'Ridge' e 'Lasso' que minimizam a função objetivo L(**w**) = 1/(2N) RSS(**w**) + \lambda ||**w**||^q. Apresente resultados para um \lambda = 0.25;

(e) Aplicando as regressões 'Ridge' e 'Lasso' e utilizando k-fold cross-validation, é possível selecionar um valor para \lambda que resulta em um modelo com melhor capacidade de generalização. Isso é feito selecionando o \lambda relativo à menor estimativa do erro de predição quadrático médio (usualmente chamado de *validation score*) ao longo dos k-folds. Também é possível selecionar um valor de \lambda que seleciona o modelo mais simples dentro de uma tolerância da estimativa do erro de predição quadrático médio. Isso é particularmente útil quando se

deseja encontrar soluções esparsas (no caso do Lasso) ou de menor norma L2 (no caso do Ridge). Para tal, um critério comumente adotado é a 'Regra de 1 desvio padrão', onde escolhe-se o maior \lambda cujo *validation score* seja igual ou pouco menor do que o 'score mínimo'+ '1 desvio padrão do score mínimo' .
- Monte as curvas de *validation score* de k-fold cross validation em função de \lambda para os modelos regularizados por 'ridge' e 'lasso'; (Sugestão: use k = 10, e procure \lambda em um intervalo [0, 0.5]);
- Calcule o desvio padrão do 'score' mínimo em cada respectiva curva e desenhe-o como barra de erro em torno daquele ponto;
- Determine o \lambda que resulta no modelo mais simples de acordo com a 'Regra de 1 desvio padrão'.
- Treine o modelo final 'ridge' e 'lasso' utilizando todos os dados (de treinamento) e o respectivo \lambda encontrado e apresente os resultados;

(f) Utilizando o conjunto de teste construído no item (b), calcule a estimativa do erro de predição quadrático médio do conjunto de teste para cada modelo (mínimos quadrados, 'ridge' e 'lasso'). Disserte sobre os resultados obtidos.

(Bônus) Estime o desvio padrão dos coeficientes do modelo obtido pelo método de bootstrap dos resíduos;
Dica: Veja os slides 9 e 10 de http://www.est.ufmg.br/~cristianocs/MetComput/Aula8.pdf