Exercises

MALIS - Fall 2019

January 3, 2020

1 Linear Models

Exercise 1: Bishop. Solve the following exercises from Bishop's book: 3.3, 4.8, 4.9, 4.10, 4.12, 4.13, 4.14, 4.15, 4.16, 4.17, 4.18, 4.19

Exercise 2: Elements of Statistical Learning. Solve the following exercises from Hastie's book: 2.1, 2.2, 2.3, 2.4, 3.2, 4.2, 4.4.

Exercise 3: The statistical properties of linear regression estimates. Using the definition of expected value and the covariance, obtain estimates for the following:

1. $\mathbb{E}[\hat{\mathbf{w}}]$
2. $cov[\hat{\mathbf{w}}]$
3. $\mathbb{E}[\hat{\sigma}]$

Based on these results, what can you say about the least squared estimator?

Exercise 4: Sigmoid function. Show that the logistic sigmoid function satisfies the property $\sigma(-a) = 1 - \sigma(a)$ and that its inverse is given by $\sigma - 1(y) = \ln y/(1-y)$ (Bishop 4.7).

Exercise 5: Discriminant analysis. Let $(X, Y) \in \mathbb{R}^d \times 0, 1$ be a random pair such that $P(Y = k) = \pi_k > 0(\pi_0 + \pi_1 = 1)$ and the conditional distribution of X given Y is $X|Y \sim \mathcal{N}(\mu, \Sigma)$, where $\mu_0 \neq \mu_1 \in \mathbb{R}^d$ and $\Sigma_0, \Sigma_1 \in \mathbb{R}^{d\times d}$ are mean vectors and covariance matrices respectively.

1. What is the (unconditional) density of X?
2. Assume that $\Sigma_0 = \Sigma_1 = \Sigma$ is a positive definite matrix. Compute the Bayes classifier h^* as a function of $\mu_0, \mu_1, \pi_0, \pi_1$ and Σ. What is the nature of the sets $h^* = 0$ and $h^* = 1$?
3. Assume now that $\Sigma_0 \neq \Sigma_1$ are two positive definite matrices. What is the nature of the sets $h^* = 0$ and $h^* = 1$?

Exercise 6: QDA. Write a computer program to perform a quadratic discriminant analysis by fitting a separate Gaussian model per class. Try it out on the vowel data, and compute the misclassification error for the test data. The data can be found in the book website www-stat.stanford.edu/ElemStatLearn. **Do not use any third party library for the QDA.** (ELS 4.9).

2 Optimization

Exercise 1. Construct an example showing that the 0-1 loss function may suffer from local minima; namely, construct a training sample $S \in (X \times \pm 1)^m$ (say, for $X = \mathbb{R}^2$), for which there exist a vector w and some $\epsilon > 0$ such that:

1. For any $\mathbf{w}'$ such that $||w - w'|| \leq \epsilon$ we have $L_S(\mathbf{w}) \leq L_S(\mathbf{w}')$ (where the loss here is the 0-1 loss). This means that $\mathbf{w}$ is a local minimum of L_S (loss over S).
2. There exists some $\mathbf{w}^*$ such that $L_S(w^*) < L_S(w)$. This means that w is not a global minimum of L_S.

(from Shalev-Shwartz and Ben-David)

Exercise 2. For at least two of the following papers identify the function that is being optimized:
https://arxiv.org/pdf/1505.03540.pdf
https://icml.cc/2012/papers/674.pdf
https://apps.dtic.mil/dtic/tr/fulltext/u2/a513243.pdf
http://cis.csuohio.edu/∼sschung/CIS660/DeepFaceRecognition_parkhi15.pdf

3 Neural networks

Exercise 1: Bishop. Solve the following exercises from Bishop's book: 5.4, 5.5, 5.6, 6.7, 5.8, 5.9, 5.10

Exercise 2: Elements of Statistical Learning. Solve the following exercises from Hastie's book: 4.6, 11.2, 11.3, 11.4.

Exercise 3: Perceptron Suppose we modify the Perceptron algorithm as follows: In the update step, instead of performing $w^{(t+1)} = w^{(t)} + y_i x_i$ whenever we make a mistake, we perform $w^{(t+1)} = w^{(t)} + \alpha y_i x_i$ for some $\alpha > 0$. Prove that the modified Perceptron will perform the same number of iterations as the vanilla Perceptron and will converge to a vector that points to the same direction as the output of the vanilla Perceptron (from Shalev-Shwartz and Ben-David).

4 ML Topics

Exercise 1. Consider the following toy problem:

$$Y \sim \mathcal{N}(\beta^*, 1)$$

where β is a real-valued parameter ($d = 1$).

1. Find the three estimators when minimizing the following three functions:
   a) $\frac{1}{2}(Y - \beta)^2 + \lambda$
   b) $\frac{1}{2}(Y - \beta)^2 + \lambda|\beta|$
   c) $\frac{1}{2}(Y - \beta)^2 + \lambda\beta^2$
2. Show a plot of the estimators as functions of the unconstrained least squares estimator (LSE) and explain the use of the following terminology for the penalized procedures: hard thresholding, soft thresholding and shrinkage.

Exercise 2: Failure of k-fold cross validation Consider a case in that the label is chosen at random according to $P[y = 1] = P[y = 0] = 1/2$. Consider a learning algorithm that outputs the constant predictor $h(x) = 1$ if the parity of the labels on the training set is 1 and otherwise the algorithm outputs the constant predictor $h(x) = 0$. Prove that the difference between the leave-one out estimate and the true error in such a case is always 1/2 (from Shalev-Shwartz and Ben-David).

Exercise 3. Let $\mathcal{H}_1, \ldots, \mathcal{H}_k$ be k hypothesis classes. Suppose you are given m i.i.d. training examples and you would like to learn the class $\mathcal{H} = \bigcup_{i=1}^{k} \mathcal{H}_i$. Consider two alternative approaches:

- Learn $\mathcal{H}$ on the m examples using the empirical risk minimization (ERM) rule
- Divide the m examples into a training set of size $(1 - \alpha)m$ and a validation set of size αm, for some $\alpha \in (0, 1)$. Then, apply the approach of model selection using validation. That is, first train each class $\mathcal{H}_i$ on the $(1 - \alpha)m$ training examples using the ERM rule with respect to $\mathcal{H}_i$, and let $\hat{h}_1, \ldots, \hat{h}_k$ be the resulting hypotheses. Second, apply the ERM rule with respect to the finite class $\{1\}$ on the αm validation examples.

  Describe scenarios in which the first method is better than the second and vice versa (from Shalev-Shwartz and Ben-David).

Exercise 4: Elements of Statistical Learning. Solve the following exercises from Hastie's book: 2.9, 3,5, 3.6, 3.12, 7.1, 7,4

Exercise 5: Ridge Regression. Show that ridge regression is biased as $\mathbb{E}[\hat{\mathbf{w}}] \neq \mathbf{w}$

5 Support Vector Machines and Kernels

Exercise 1: Kernelized regression. An important issue in kernel methods is overfitting. When the feature space is in infinite-dimensional, it is easy to achieve perfect performance on any training set which often leads to severe overfitting. This is best seen on a kernelized formulation of ordinary least squares regression.
Derive the kernelized version of ordinary least squares and show that if the kernel is positive definite, it is indeed possible to classify any training set perfectly. An important consequence of this is that you should generally avoid using kernel methods without regularization.

Exercise 2: Soft SVM. Given the a dataset in 1-d space, which consists of 4 positive data points $(0, 1, 2, 3)$ and 3 negative data points $(-3, -2, -1)$. Suppose that we want to learn a soft-margin linear SVM for this data set. Remember that the soft margin linear SVM can be formalized as a constrained quadratic optimization problem. In this formulation, C is a regularization parameter, which controls trade-off between slack variable penalty and the margin.

- if $C = 0$ how many support vectors do we have?
- if $C = \infty$ how many support vectors do we have?

Exercise 3: Elements of Statistical Learning. 12.1, 12.2

Exercise 4: Bishop. 6.1, 6.2, 6.3, 6.10, 7.1, 7.2, 7.3, 7.4, 7.5, 7.6, 7.7

6 VC dimension

For hypothesis class $\mathcal{H}$, its VC dimension is at least d if there exists some samples $|S| = d$ which is shattered by $\mathcal{H}$. Please note that this does not mean all samples of size d need to be shattered by $\mathcal{H}$. To show that the VC dimension is at most d, one must show that no sample of size $d + 1$ could be shattered by $\mathcal{H}$. You will calculate the VC-dimension of some hypothesis classes. Remember you need to show the following 2 steps to prove the VC dimension of $\mathcal{H}$ is d:

- There exists a set of d points which can be shattered by $\mathcal{H}$;
- There exists no set of $d + 1$ points which can be shattered by $\mathcal{H}$

Exercise 1. $\mathcal{X} = \mathbb{R}$, $\mathcal{H}$ is the union of two intervals.

Exercise 2. $\mathcal{X} = \mathbb{R}^2$, $\mathcal{H}$ is the set of axis-parallel squares (with equal height and width).

Exercise 3. $\mathcal{X} = \mathbb{R}$, $\mathcal{H}$ is the set of functions defined as:

$$h(x) = sign(sin(ax + b))$$

where a and b are parameters in h, $sign(x) = +1$ if $x \geq 0$ and $sign(x) = -1$ otherwise.

7 Trees & Ensembles

Exercise 1: Trees Suppose a data set with M binary input features and R training set samples. What is the maximum possible number of leaves of a decision tree trained on this data set?

Exercise 2: Bagging Show that any bootstrap sample D' follows the original distribution P from the complete training set D.

Exercise 3: Bagging, Bootstrap and OOE What is the probability that a sample is never chosen N times for a set of N observations?

Exercise 4: Elements of Statistical Learning. 10.1, 10.2, 10.10, 15.1, 15.2