Exercises Pattern Recognition 2018
Linear Models, Optimization and Support Vector Machines

1 Linear Regression

A professor suspects that the performance of his students depends on the temperature in the exam room according to the model

$$t = w_0 + w_1 x + \varepsilon$$

Here x denotes the temperature in degrees centigrade, and t denotes the performance of a student (in some unit of measurement). The relationship is supposed to hold for $20 \leq x \leq 35$. To quantify this model, he collects the following 7 observations:

n	1	2	3	4	5	6	7
x_n	31	25	27	23	32	22	29
t_n	80	105	120	105	70	120	100

(a) Compute the least squares estimates of w_0 and w_1.

(b) Interpret the values of the estimates you have found under (a), that is, what do they mean?

(c) Use the fitted model to predict student performance when the temperature in the exam room is 20 degrees centigrade.

(d) What percentage of the variation in performance is explained in this model by the variation in temperature?

2 Linear Models for Classification

It has often been claimed that the death penalty is applied in a racially discriminatory fashion. Data were provided by the Georgia Parole Board, the Georgia Supreme Court, and lawyers involved in the cases on the following variables:

variable	description
death	1 if got death penalty; 0 otherwise
blkdef	1 if black defendant; 0 otherwise
whtvict	1 if white victim; 0 otherwise
aggcirc	number of aggravating circumstances
fevict	1 if the victim is female; 0 otherwise
stranger	1 if victim is stranger; 0 otherwise
multvict	1 if 2 or more victims; 0 otherwise
multstab	1 if multiple stabs; 0 otherwise
yngvict	1 if victim 12 years or younger; 0 otherwise

We fitted a linear regression model to this data set, with `death` as the target variable. The results are summarized below:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.18679    0.20034  -0.932 0.353609
blkdef      -0.08692    0.11024  -0.788 0.432482
whtvict      0.30522    0.12075   2.528 0.013202
aggcirc      0.06787    0.03714   1.827 0.070947
fevict       0.07903    0.10613   0.745 0.458409
stranger     0.35639    0.10146   3.512 0.000693
multvict     0.04994    0.13940   0.358 0.720987
multstab     0.28365    0.15177   1.869 0.064845
yngvict      0.05036    0.17730   0.284 0.777044

# summary of the fitted probabilities on the training data

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.1380  0.3489  0.5038  0.4900  0.6859  0.9320
```

(a) Which explanatory variables have a coefficient that is significantly different from zero at significance level $\alpha = 0.05$? And at $\alpha = 0.1$?

(b) According to this model, what is the probability that the defendant gets the death penalty when he or she is black, the victim is an asian man of 40 years old, the defendant and victim were good friends, the victim was strangled, and there were no aggravating circumstances.

(c) All else equal, according to this model, what is the difference in probability of the death penalty between a case where the victim and defendant knew each other and a case where victim and defendant were strangers?

(d) Interpreting the fitted coefficients and their p-values, would you say there is any evidence of racial discrimination in the application of the death penalty? Explain.

We also fitted a logistic regression model to the same data set.
The results are summarized below:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.5675     1.1243  -3.173 0.001508
blkdef       -0.5308     0.5439  -0.976 0.329059
whtvict       1.5563     0.6161   2.526 0.011528
aggcirc       0.3730     0.1963   1.900 0.057447
fevict        0.3707     0.5405   0.686 0.492829
stranger      1.7911     0.5386   3.325 0.000883
multvict      0.1999     0.7450   0.268 0.788490
multstab      1.4429     0.7938   1.818 0.069082
yngvict       0.1232     0.9526   0.129 0.897132

# summary of the fitted probabilities on the training data

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03374 0.30220 0.48820 0.49000 0.71040 0.90180
```

(e) According to this model, what is the probability that the defendant gets the death penalty when the conditions are the same as under (b)?

(f) Interpretation of the coefficients and the marginal effect of variables on the outcome is a bit more difficult than in the linear probability model. The fitted response function is given by

$$\hat{p}(t = 1|\boldsymbol{x}) = (1 + e^{-\boldsymbol{w}_{\text{ML}}^{\top}\boldsymbol{x}})^{-1},$$

where $\boldsymbol{w}_{\text{ML}}$ are the maximum likelihood estimates of the coefficients $\boldsymbol{w} = [w_0, w_1, \ldots, w_{M-1}]^{\top}$. To assess the marginal effect of an increase in x_i on the fitted probability for class 1, determine:

$$\frac{\partial \hat{p}(t = 1|\boldsymbol{x})}{\partial x_i},$$

where x_i is the i-th predictor variable, not the i-th observation of x.

3 Logistic Regression

In a study of commuting, for 21 persons their travel time to work by car and by public transport is determined. Also, each person in the study is asked whether he or she actually travels to work by car or public transport. Using these data, we estimate the model

$$p(t_n = 1 \mid x_n) = \frac{\exp(w_0 + w_1 x_n)}{1 + \exp(w_0 + w_1 x_n)},$$

where $t_n = 1$ means that person n travels to work by car, $t_n = 0$ that person n travels by public transport, and $x_n =$ (travel time by public transport $-$ travel time by car) for person n (in minutes). This produces the following maximum likelihood estimates

$$w_0 = -0.24 \qquad w_1 = 0.053$$

(a) We note that w_1 has a positive sign. Is this surprising? Explain.

(b) We also note that w_0 has a negative sign. Give a simple interpretation of this finding.

(c) According to this model, what is the probability that someone travels to work by car, if public transport takes 30 minutes longer?

(d) What is the marginal effect on the probability of choosing to travel by car, of an increase in x at $x = 5$? And at $x = 30$?

(e) Use the fitted model to give a simple classification rule for new cases.

4 Optimization/Linear Regression

Just for practice, let's solve a linear regression problem from first principles, that is, without using the formulas we derived for w_0 and w_1.

We are given the following three observations on x and t:

n	x_n	t_n
1	1	4
2	2	8
3	3	6

We want to fit a linear regression model

$$y(x) = w_0 + w_1 x$$

by the method of least-squares.

(a) Specify the sum of squared errors function $E(w_0, w_1)$ for this specific data set.

(b) Determine the partial derivatives $\frac{\partial E}{\partial w_0}$ and $\frac{\partial E}{\partial w_1}$ for the error function you found under (a). Equate both partial derivatives to zero and solve for w_0 and w_1.

(c) Determine the second order partial derivatives, and put them in the Hessian matrix. Verify that we have indeed found a minimum (rather than maximum or saddle point) by ascertaining that the Hessian matrix is positive definite for the values of w_0 and w_1 that you found under (b).

5 Optimization/Linear Regression

In simple linear regression, the sum of squared errors is given by:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(t_n - w_0 - w_1 x_n)^2,$$

where $\mathbf{w} = [w_0 \;\; w_1]^\top$ denotes the weight vector to be estimated from the data.
Suppose we want to minimize this error function using the method of gradient descent.

(a) Derive an expression for the gradient $\nabla E(\mathbf{w})$.

(b) Let $\mathbf{w}^{(0)} = [1.6 \;\; 0.8]^\top$, and the step size (learning rate) $\eta = 0.1$.
Use the single data point $t_n = 3, x_n = 3$ to update the weight vector.

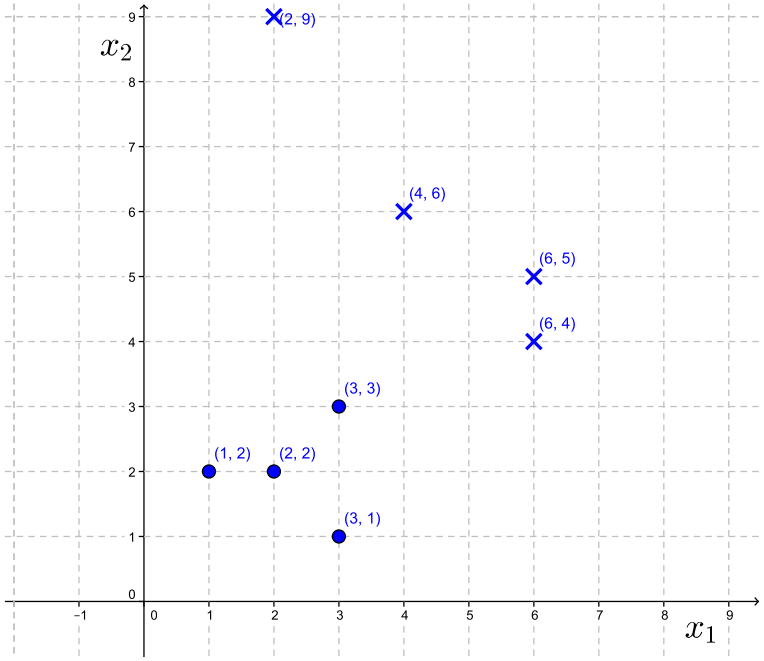
(c) Verify whether the update has decreased the squared prediction error for the data point used. What if η were 0.2 instead of 0.1?

6 Support Vector Machines

We receive the following output from the optimization software for fitting a support vector machine with linear kernel and perfect separation of the training data:

n	$x_{n,1}$	$x_{n,2}$	t_n	a_n
1	1	2	-1	0
2	2	2	-1	0
3	3	1	-1	0
4	3	3	-1	$\frac{1}{4}$
5	2	9	$+1$	0
6	4	6	$+1$	$\frac{1}{8}$
7	6	4	$+1$	$\frac{1}{8}$
8	6	5	$+1$	0

Here $x_{n,1}$ denotes the value of x_1 for the n-th observation, a_n is the value of the Lagrange multiplier for the n-th observation, etc. The figure below is a plot of the same data set, where the dots represent points with class -1, and the crosses points with class $+1$.



You are given the following formulas:

$$b = t_m - \sum_{n=1}^{N} a_n t_n \mathbf{x}_m^\top \mathbf{x}_n \qquad \text{(for any support vector } \mathbf{x}_m\text{)}$$

$$y(\mathbf{x}) = b + \sum_{n=1}^{N} a_n t_n \mathbf{x}^\top \mathbf{x}_n$$

Answer the following questions:

(a) Give the support vectors for this problem.

(b) Compute the value of the SVM bias term b.

(c) Which class does the SVM predict for the data point $x_1 = 0, x_2 = 7$?

(d) Give the equation $y = b + \mathbf{w}^\top \mathbf{x}$ of the maximum margin decision boundary. Draw it in the graph.