Introduction to Statistical Learning
Revision Quiz Answers

Department of Mathematics
Imperial College, London

Question

Correct answers are highlighted in BLUE.

Question 1

Let $\hat{\beta}$ be the least-squares estimator for the general multivariate linear regression model. Then $\hat{\beta}$

(a) has zero mean;

(b) is unbiased and has the smallest variance amongst all linear unbiased estimators;

(c) always has the smallest mean squared error;

(d) is shrunk towards zero.

Question 2

The ridge regression problem in the usual linear regression setup is

$$\hat{\beta}_R(\lambda) = \text{argmin}_\beta \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} X_{i,j}\beta_j \right)^2 + \lambda \beta^T \beta \right\}. \quad (1)$$

When λ is extremely large, what is $\hat{\beta}_R(\lambda)$?

(a) slightly shrunk least squares estimates;

(b) $\hat{\beta}_0 = \bar{Y}$ and $\hat{\beta}_j = 0$ for $j = 1, \ldots, p$;

(c) the penalty is ignored and $\hat{\beta}_R$ become the least squares estimates;

(d) $\beta_0 = \bar{Y}$ and $\beta_j = 0$ for $j = 1, \ldots, p$.

Question 3

In what circumstances would one consider using ridge regression?

(a) when you think one explanatory variable is more important than the others;

(b) when the model error is Gaussian;

(c) when the response is not a linear function of the explanatory variable;

(d) when the explanatory variables are all highly correlated.

Question 4

In the Lasso, for the orthogonal design, the optimal coefficients can be obtained from the least squares ones, $\hat{\beta}_{LS}$, by

(a) $\hat{\beta}_{LS} - \lambda$;

(b) $\hat{\beta}_{LS}\, \mathbb{I}(|\hat{\beta}_{LS}| > \lambda)$;

(c) $\text{sgn}(\hat{\beta}_{LS})(|\hat{\beta}_{LS}| - \lambda)^{+}$;

(d) $\text{sgn}(\hat{\beta}_{LS})\, \mathbb{I}(|\hat{\beta}_{LS}| > \lambda)$.

Question 5

Principal components regression and ridge regression can been seen as modifiers of the eigenvalues/singular values of an associated eigen/singular value decomposition. The modification for principal components is:

(a) setting to zero coefficients whose eigenvalues are small;

(b) removing coefficients whose eigenvalues are smaller than the entropy;

(c) shrinking the coefficients, shrinking more for those associated with small eigenvalues;

(d) soft-thresholding the coefficients associated with small eigenvalues.

Question 6

Which country is hosting this year's Olympics?

(a) China;

(b) USA;

(c) Brazil;

(d) Japan.

Question 7

If X is a $n \times p$ data matrix, let B and E be its associated inner product and Euclidean distance matrices, respectively.
Which of the following is *incorrect*?

(a) $e_{m,\ell} = 2b_{m,m} + 2b_{\ell,\ell} - b_{\ell,m}$;

(b) $e_{m,\ell} = b_{m,m} + b_{\ell,\ell} - 2b_{\ell,m}$;

(c) $e_{m,\ell} = \sum_{j=1}^{p}(X_{m,j} - X_{\ell,j})^2$;

(d) $b_{m,\ell} = -\frac{1}{2}(e_{m,\ell} - \frac{e_{m,\bullet}}{n} - \frac{e_{\bullet,\ell}}{n} + \frac{e_{\bullet,\bullet}}{n^2})$, where $\bullet$ indicates summation over the index.

Question 8

In classical scaling, a configuration is obtained from the eigendecomposition of the inner product matrix B.
There is always one zero eigenvalue. If there are non-trivial negative eigenvalues, then they are indicators of:

(a) the Euclidean distance matrix being too big;

(b) the configuration not being properly centred;

(c) non-Euclidean nature of the original distances/dissimilarities;

(d) numerical error arising from the numerical linear algebra routines.

Question 9

Given two sites, x, y, which are being compared on a list of attributes, let the counts of attributes, a, b, c, d, in the four different presence/absence situations be:

		x attributes	
		Yes	No
y attributes	Yes	a	b
	No	c	d

and $a, b, c, d \geq 0$. The Jaccard distance is given by:

(a) $(b + c)/(a + b + c)$;

(b) $(a + d)/(a + b + c + d)$;

(c) $(b + c)/(a + b + c + d)$;

(d) $d/(a + b + c)$.

Question 10

Given a multivariate data set, X, the k-means clustering algorithm

(a) works out the best projection of the data based on the known class membership;

(b) is identical to the self-organizing map;

(c) is a grouping of individuals into groups, where the points in the group are closest to the mean of that group;

(d) automatically works out the number of clusters in the data set.

Question 11

The stress in ordinal scaling measures:

(a) the distance between the set of distances $\{d_{m,\ell}\}$ created from a known trial configuration and the ordered dissimilarities $\{\delta_{m,\ell}\}$ computed from the data, divided by a normalisation energy based computed from the $\{d_{m,\ell}\}$.

(b) the distance between the set of dissimilarities $\{\delta_{m,\ell}\}$ created from a known trial configuration and the least-squares monotone regression fit $\hat{\delta}_{m,\ell}$ of the $\delta_{m,\ell}$ to ordered distances $\{d_{m,\ell}\}$ computed from the data, divided by a normalisation energy based computed from the $\{d_{m,\ell}\}$.

(c) the distance between the set of distances $\{d_{m,\ell}\}$ created from a known trial configuration and the least-squares monotone regression fit $\hat{d}_{m,\ell}$ of the $d_{m,\ell}$ to ordered dissimilarities $\{\delta_{m,\ell}\}$ computed from the data, divided by a normalisation energy based computed from the $\{d_{m,\ell}\}$.

Question 12

The following is a list of 'twin cities' (or 'sister cities') of New York City. Which one has been a twin for longest?

Beijing, Budapest, Cairo, Brasilia, Jerusalem, Baku, Johannesburg, London, Madrid, Rome, Santo Domingo, Tokyo.

Question 13

In spline smoothing, the effective degrees of freedom is equal to the trace of the smoother matrix with parameter λ.
If λ is extremely high, then effective degrees of freedom is close to:

(a) three as the spline is a cubic spline;

(b) two;

(c) the number of observations plus two (for the free degrees of freedom at the ends of the spline);

(d) zero.

Question 14

Which of the following is not a typical property of a kernel function, $K(x)$, used for kernel density estimation:

(a) $K(x) \geq 0$;

(b) $\int K(x)\, dx = 1$;

(c) $K(x) = K(-x)$;

(d) $K(0) = 1/\sqrt{2\pi}$.

Question 15

In local linear regression, generally speaking, which has the best bias properties:

(a) the Nadaraya-Watson estimator;

(b) the local linear estimator;

(c) local quadratic regression.

Question 16

Assume an additive signal plus noise model, with independent and identically distributed Gaussian noise. In wavelet shrinkage, hard thresholding is often use to modify noisy coefficients to produce an estimator. This is because:

(a) the expected mean-squared error can be proven to be smallest with hard thresholding;

(b) it is quick to compute;

(c) the signal gets squeezed into few significant coefficients, with many zeroes, but their overall energy is the same as the input signal and the noise affects all coefficients in the same way (independently);

(d) because it doesn't make sense to take the mean of coefficients with different variances.

Question 17

When carrying out kernel regression, the mean-squared error is minimised when the bandwidth h is

(a) large;

(b) chosen to balance the bias and variance;

(c) very small;

(d) equal to $1.06\hat{\sigma}n^{-1/5}$.

Question 18

In what year did the *Normal School of Science* turn into the *Royal College of Science*?
(These were precursors of Imperial College).

(a) 1907;

(b) 1851;

(c) 1890;

(d) 2002 .

Question 19

Exploratory projection pursuit is a technique that

(a) projects multivariate data into low dimensions and searches for projections that maximise the closeness to normality in the projected data's density estimate;

(b) usually more computationally efficient than principal components analysis;

(c) guaranteed to find more structure compared to principal components analysis;

(d) projects multivariate data into low dimensions and searches for projections that look for departures from normality in the projected data's density estimate.

Question 20

Factor analysis tends to be unpopular because:

(a) factors can be rotated within their own space to obtain equivalent, but different, factors;

(b) factor analysis cannot work with large data sets;

(c) factor analysis needs independent variable to work on;

(d) it does not give answers that are much different from principal components analysis.

Question 21

Let $H(Y)$ be the entropy of random variable Y, which might be a p-dimensional vector. Let $Y = A^T X$, where X has identity covariance matrix and A orthgonal.
Then the independent components analysis objective function, $I(Y)$, is

(a) $\sum_{j=1}^{p} H(Y_j) - H(X)$;

(b) $\sum_{j=1}^{p} \{H(Y_j) - H(X_j)\}$;

(c) $\sum_{j=1}^{p} \{H(Y_j) - H(X_j)\}^2$;

(d) $\sum_{j=1}^{p} Y_j \log(Y_j)$.

Question 22

Why is the single layer perceptron similar to projection pursuit regression, when the weights are small?

(a) the back-propagation algorithm is easier to calculate as some terms can be ignored;

(b) the learning rate is orthogonal to the weights, which has a direct analogue in PPR;

(c) because the activation function is effectively linear in the range of small parameters;

(d) many of the weights will be very close to zero, resulting in a sparse current solution, which means fewer nodes are involved, hence mimicing the PPR solution.

Question 23

What is the difference between scaling variables and whitening?

(a) scaled variables are usually correlated, whitened ones are not;

(b) scaled variables have zero mean, whitened have not;

(c) whitened variables are fixed to be on the range of $[-1, 1]$, scaled ones are not;

(d) whitening can only be applied to Gaussian random variables, but scaling can be computed on variables of any distribution.

Question 24

Put these fast-food restaurants in order, from largest to smallest (in terms of numbers of locations):

Burger King, KFC, McDonald's, Starbucks, Subway.

Subway, McDonald's, Starbucks, KFC, Burger King

Question 25

Which of the following is a widely understood weakness of classification and regression trees:

1. slow to compute;
2. trees are hard to explain to others;
3. they lack continuity, small changes in the input variables can result in completely different trees;
4. a single estimated tree usually gives poor results.