STATG006: Solutions to Exercise Sheet #8

The exercises in this sheet focus on assorted questions on unsupervised learning. As before, we provide solutions, sometimes detailed, sometimes a sketch that should point you to the complete solution. Sketches should not be taken at face value as the level of detail required for an exam answer.

1. One example is a distribution we have mentioned in our lectures in linear regression: say X follows a distribution that puts more weight on its tails compared to a Gaussian with same mean and variance, like the double exponential distribution (also known as the Laplace distribution). The $P(X > c)$ is actually higher under the true model instead of the assumed Gaussian with the same mean and variance. If we are interested in knowing whether a particular tail event is likely, we will underestimate this probability if we assume Gaussianity even if we get the correct estimates of mean and variance.

2. Please run the corresponding script in EX8.R. Basically, the observations we can make are: (a)+(b)+(c) Eyeballing the right amount of smoothing can be very difficult. The results picked by cross-validation do not need to correspond to our intuition. Even with confidence bands, we cannot say that the difference between our intuition and the LOOCV choice are due to sampling variability of our training set. (d) The complexity in the distribution of the refractive index is not due only to pre-defined classes (which are in one sense man-induced, as the choice of glass material in the different classes is not a feature of Nature itself, but an engineering choice). The distribution within each class (taken into consideration that confidence bands here may be a bit wide) are themselves far from Gaussian, where even the modes of the inferred nonparametric and the Gaussian fit can have a big mismatch. Finally, classification will be affected by class boundaries of the densities. Comparing some of them highlight that the Gaussianity assumption is still harmful even in this easier task of classification.

3. For part (a): essentially this works in two stages. First, we get an estimate of the cdf of the variable. Let's call it $\hat{F}_n(x)$. We know that if X follows a continuous distribution with cdf $F(x)$, then $F(X)$ follows the uniform distribution in the interval $[0, 1]$ (this was mentioned in the slides of Chapter 2). We also know that if some variable U follows an uniform in $[0, 1]$, then $G^{-1}(U)$ follows the distribution given by the cdf $G(\cdot)$:

$$\begin{aligned} P(U \leq u) &= u \text{ (because } U \text{ follows the uniform in } [0,1]) \\ P(U \leq u) &= P(G^{-1}(U) \leq G^{-1}(u)) \\ &\Rightarrow P(V \leq q_u^G) = u, \end{aligned}$$

where $V \equiv G^{-1}(U)$ and $q_u^G \equiv G^{-1}(u)$, by definition the u-th quantile of the cdf $G(\cdot)$. This holds for all $u \in [0,1]$. Hence, the cdf of V has to be $G(\cdot)$.

The `F_tilde` array in GAUSSIFY is the estimate of $F(x)$ we use. It is basically the empirical cdf but for the points close to the smallest and largest values of our sample. Instead, for empirical values less than $\delta \equiv 1/(4n^{0.25}\sqrt{\pi \log(n)})$, we threshold it to be δ, with a similar idea for values greater than $1-\delta$ (this choice of δ is justified by some theory explained in the given reference). The reason for that is that the empirical cdf estimate has high variance at the endpoints, and as a matter of fact its inverse will be infinite at the largest sample value (since $\hat{F}_n(x_{max}) = 1$ for x_{max} the maximum of the sample, and $F^{-1)}(1) = \infty$ for unbounded distributions).

Once we agree on the estimate of $F(x)$, we can get uniformly distributed variables $V^{(i)} \equiv \hat{F}(X^{(i)})$. Following the reasoning above where G is the cdf of standard Gaussian, we can get values $z^{(i)} \equiv \Phi^{-1}(v^{(i)})$, which will then theoretically follow a standard Gaussian.

For (b): please see the solution in EX8.R.

4. (a) As both X_1 and Y_1 are standard Gaussians, then these events will have exactly the same probability of happening.

   (b)
$$P(0 \leq X_1 \leq 3 \text{ and } 0 \leq X_2 \leq 3) =$$
$$\int_0^3 \int_0^3 \frac{1}{2\pi} \times \begin{vmatrix} 1 & 0.4 \\ 0.4 & 1 \end{vmatrix}^{-1/2} \exp\left\{ -\frac{1}{2}[x_1 \ x_2] \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right\} dx_1 dx_2$$

   (c) In principle we need to do the computations, but here we have a special case that facilitates knowing the result without having to do any calculation: in the case of (Y_1, Y_2), we know that the contours of the distribution are circles centered at zero. If we imagine the circle of radius 3, $P(0 \leq Y_1 \leq 3 \text{ and } 0 \leq Y_2 \leq 3)$ will put mass equally on all four quadrants given by the $Y_1 = 0$ and $Y_2 = 0$ axes. In contrast, the ellipse given by the joint distribution of (X_1, X_2) will put more mass in the upper right and bottom left than in the upper left and bottom right. So the upper right quadrant will have more mass in the (X_1, X_2) case than in the (Y_1, Y_2) case.

   If we wanted to check this numerically, the following R code solves the integral of part (b) for both cases:

```
> library(mvtnorm) # You may need to install this package
> pmvnorm(lower = c(0, 0), upper = c(3, 3), mean = c(0, 0),
+ corr = matrix(c(1, 0.4, 0.4, 1), ncol = 2)) # X1, X2
[1] 0.3130494
> pmvnorm(lower = c(0, 0), upper = c(3, 3), mean = c(0, 0),
+ corr = matrix(c(1, 0, 0, 1), ncol = 2)) # Y1, Y2
[1] 0.2486519
```

5. (a) Although we can show this formally, let us state without formal proof that if Y and X_i are uncorrelated given the remaining covariates, then the corresponding regression coefficient $\beta_i^\star$ will be zero.

   (b) If X_i and X_j are strongly correlated given the other covariates, then their corresponding regression coefficients may be close to zero not necessarily because X_i and X_j are not important to predict Y, but because each individually "masks" the other (Y will be weakly correlated with X_i given X_j if X_i is "mostly determined" by X_j). In practice, this may translate into instabilities while doing regression, as explained in Chapter 3 when we talked about collinearity.

   (c) If all covariates are mutually independent, then consider the 'partial covariance" of Y and X_1, $E[YX_1 \mid x_2, \dots, x_p]$. By mutual independence, this is given by
$$E[\beta_1^\star X_1^2 + \beta_2^\star X_1 x_2 + \cdots + \beta_p^\star X_1 x_p + X_1 \epsilon] = \beta_1^\star \sigma_1^2,$$
   where σ_1^2 is the variance of X_1. So, at least in the population, this is equivalent to just regressing Y on X_1. We can do "marginal regressions" of Y on each X_i and use a decision rule (like hypothesis tests with Bonferroni corrections) to decide which coefficients are zero without ever doing a combinatorial search.

6. This question was intended to be solved by hand, so you can have a concrete grasp of the K-means clustering algorithm. Instead of showing hand-made calculations, I provide a solution in EX8.R.

7. In the first case, buying/not buying a computer gives "one unit" of discrimination between individuals, which is of a much smaller scale than the buying of socks. K-means here will roughly divide shoppers by those buying (approximately) below the average and above the average number of socks.

   In the second case, buying a computer will have about the same effect as buying as many socks as bought by the largest sock purchase in the data. This will give a partition of the shoppers that have a purchasing activity that can include customers with large and small number of socks bought in a same group, as long as those with low-sock behaviour compensate by having bough a computer (a more precise answer would require knowing the joint behaviour of activity, instead of the marginal ones).

   In the third case, it should be obvious that the influence of sock has has been nearly wiped out and we will get are a group of 3 customers who did not buy a computer in one cluster, and four which bought a computer in the second cluster.

8. (a) The total variance of a sample is given by the sum of the squares of every number in the data, and it is meant to capture the "spread" of information in the sample (it is equivalent to the sum of empirical variances over all variables, assuming zero empirical mean). We can define that also for the data projected into the first principal component. "10/projected sample, compared to the total variance of the original sample. This low number seems to indicate that

some of the signal dependency across tissues is captured by the first component, although its source is unclear.

(b) This tacitly assuming that the principal component is the contribution of the type of machine to the dependency of the tissue measurements. It may or may not be, and it can include the effect of other hidden causes that were aggregated as a latent factor (recall the interpretation of PCA as a latent variable model). Since we *know* which tissues were measured by which machines, it seems silly to ignore this information. We can split the data by machine, and do tests that compare treatment and control within each machine.

(c) Simulated data can be generated by adding 4 random variables, one for each machine and condition, to be added to the distribution of each gene. Combinations which add up to the same mean should be detected as such, but I will leave the details of this open-ended question to you.

9. For simplicity, let's say we have $K = 2$ and the number p of variables is 1. The log-likelihood of the mixture of Gaussians with common variance σ^2 can be written as

$$l(\theta_1, \theta_2, \mu_1, \mu_2, \sigma^2) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{2} p_{N(\mu_k, \sigma^2)}(y^{(i)} \mid x^{(i)} = k) P(x^{(i)} = k) \right),$$

where $\theta_k \equiv P(x^{(i)} = k)$ and $p_{N(\mu_k,\sigma^2)}$ is the pdf of a Gaussian with mean μ_k and variance σ^2. In maximum likelihood, we optimise $\theta_1, \theta_2, \mu_1, \mu_2$ and σ^2.

K-means follows from this with the follow modification: we optimize each $x^{(i)}$ instead of θ_1, θ_2. So the function to be optimized is

$$S(x^{(1)}, \dots, x^{(n)}, \mu_1, \mu_2, \sigma^2) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{2} p_{N(\mu_k, \sigma^2)}(y^{(i)} \mid x^{(i)} = k) I(x^{(i)} = k) \right),$$

where $x^{(i)} \in \{1, 2\}$ for all i, and $I(x = k)$ is the indicator function that returns 1 if $x = k$ and 0 otherwise. This is also hard to optimise, because $x^{(i)}$ is discrete (for instance, we cannot calculate gradients). But notice that the above is the same as

$$S(x^{(1)}, \dots, x^{(n)}, \mu_1, \mu_2, \sigma^2) = -\frac{n \log(\sigma^2)}{2} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(\sum_{k=1}^{2} (y^{(i)} - \mu_k)^2 I(x^{(i)} = k) \right).$$

So, if we fix μ_1 and μ_2 and optimise the above with respect to $x^{(1)}, \dots, x^{(n)}$, it should be clear that $x^{(1)}$ is given by the k that minimises $(y^{(1)} - \mu_k)^2$, $x^{(2)}$ is given by the k that minimises $(y^{(2)} - \mu_k)^2$ and so on. If we fix $x^{(1)}, \dots, x^{(p)}$, it is hopefully clear that μ_k is optimised by the average of all $y^{(i)}$ in which $x^{(i)} = k$. Parameter σ^2 is not necessary to infer μ_1, μ_2 or the assignments $x^{(i)}$.

10. See EX8.R.
11. See EX8.R.
12. See EX8.R.
13. For bragging rights, post your score in the forums!