Solutions to exercises: week 6

Exercise 6.1
75,287,520

Exercise 6.2
$\sum(Ax_i - A\bar{x})(Ax_i - A\bar{x})^T = A(\sum(x_i - \bar{x})(x_i - \bar{x})^T)A^T = ACA^T$

Exercise 6.3
(a) The easiest seems to be to start from the righthand side and to try to simplify stuff to the lefthand side. First expand $(x - x')(x - x')^T$ and then switch to the actual definition of expectation (of x *and* x'). Using severe abuse of notation, we get $\int\int[xx^T - xx'^T - x'x^T + x'x'^T] = \int xx^T + \mu_i\mu_i^T + \mu_i\mu_i^T + \int x'x'^T$. With the final remark that the first term equals the latter, we are there (or close enough).
(b) Hmpf. The actual proof might still not be easy :) Using the (by now) fact from (a), we can write out all three matrices purely in terms of samples, so without the need to refer to means etc. Once we succeeded in doing so, we readily see that indeed $S_m = S_b + S_w$. No really...

Exercise 6.4
(a) As the M means are in an $M-1$-dimensional subspace, one can construct $D - M + 1$ (with D the feature dimensionality) vectors v that are orthogonal to the subspace. Taking any such vector, one can show that $S_b v = 0$ by using the definition of S_b and realizing that $(\mu_0 - \mu_i)^T v = 0$ for every class mean μ_i.
(b) Either the space is smaller than $M - 1$ dimensions ($D < M - 1$) or the means are, in fact, in a lower-dimensional subspace (which could be considered a rather non-generic situation in case of continuous data).

Exercise 6.6
(a) Take the two points coming from the same class and draw a line through them. The 1D subspace is given by any line orthogonal to this initial line going through the two points.
(b) Take a line perpendicular to the plane through the three points from the same class.
(c) something with the shortest line connecting the two lines going through the points from the same class?
(d) When all the points of a class are projected onto one point, the between-scatter becomes zero, and the Fisher criterion becomes infinity. If we have one point per class, this is (almost) always the case, and we can make an infinite number of optimal Fisher subspaces.

Exercise 6.7
(a) Write out the variance for the projected values $x^T v$ and, at the same time, write out C and from there $v^T C v$. Finally, make sure that they turn out to be equal :)
(b) First we set up the optimisation problem. We want to maximise $L = v^T CV$ such that $v^T v = 1$. Using Langrange multipliers we get $L = v^T Cv - \alpha(v^T v - 1)$. Setting the derivative of L w.r.t. v to zero gives: $\frac{\partial L}{\partial v} = v^T C\mathbb{I} + \mathbb{I}^T Cv - \alpha v - \alpha v = 2Cv - 2\alpha v = 0$. So we need to solve the eigenvalue problem $Cv = \alpha v$.
Then to show that we need the largest eigenvalue, do the trick of multiplying $Cv = \alpha v$ both sides with v^T, and we see that $v^T CV = v^T \alpha v = \alpha \cdot 1$ is the total variance in the direction of v. We want to maximise this, so we need the largest eigenvalue.

Exercise 6.9
(a) This should be a vector pointing in the direction of the third dimension.
(b) The total covariance is given by the sum of the 3D covariance of the Gaussian plus the identity matrix, which gives diag(100,100,125) as the solution.
(c) One can do the explicit calculations and find the Eigenvector with the largest Eigenvalue, but one can also see that, as the mean within class is spherical, we will find the same component as the one under a.
(d) One should work out the product $T'CT$ which leads to the asked for solution.

(e) Considering the off diagonals, they are positively correlated.
(f) Realize that all direction along the diagonals have the same variance and that the first two features are positively correlated. This means, the largest for these two features is in the direction (1,1). Then realize that the variance in this direction must be larger than the individual variance. But this mean that in 3D, as the third feature is not correlated to the first two, that (1,1,0) has the largest variance and therefore is the first PC.
(g) inv(T)v or any variation to this will do the trick. One argument is that v "covaries" with the transformation and should, in principle, find the equivalent subspace in the transformed space. What inv(T) does is explicitly undoing the feature transform, after which one is just in the original space and can apply v, which was shown to be optimal already.

Exercise 6.10
(a) Say something about between vs. within. Precise definitions of what is a vector etc. seems necessary for a good description.
(b) Dipping. The larger the training set becomes, the closer the two means get near to each other, the less well-specified the direction becomes.
(c) Here they, again?, should say something about the means coming very close to each other. An additional remark about the within scatter, that that also doesn't help, would be good as well.
(d) Opposite behavior. Becomes better and better for classification.
(e) Second dimension does not participate. Because of equal variance in other two dimensions angle of PC should be at 45 degrees. That is, (1,0,1) is a solution.

Exercise 6.11
(a) Between covariance is $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$.
(b) Check the eigenvalues using the matrix $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.
(c) It stays exactly the same.
(d) The true error rate is typically larger.
(e) We get a decreasing curve.
(f) We get an increasing curve.
(g) We get a (lightly) decreasing curve. All main PCs are equally bad but the NMC will be slightly better the more data there is present.

Exercise 6.12
(a) One can do the explicit calculation, but is should be evident that the sought after direction is $(1, 0, 0)$ or any multiple of this vector.
(b) Because the within scatter is spherical, one should be able to see that the answer will the same as in item a.
(c) There are multiple ways to see this and , of course, one could do the explicit calculation. But generally, people should realize that the second coordinate is of least importance in PCA and that for Fisher there is no within-scatter in that same second dimension. Conclusion: PCA and Fisher should give the same solution.
(d) The answer for Fisher should be easy as it is scaling invariant, given enough samples. PCA however is not scale invariant and the scaling is just enough to make the first coordinate the least varying on. The new subspace will be equal to the second and third feature. One might have to determine the total scatter to really see this. This might be difficult for many.
(e) The within scatter will be rather unstable and therefore Fisher quicker involves the second dimension, which does not contain any discriminatory information, into the 2D representation. PCA will therefore be better in general.

Exercise 6.15
(a) Many configurations are possible of course. The easiest probably is to construct a problem where the best individual feature occurs twice and so choosing it twice won't add anything.
(b) The first feature is always the same (unless we have two features that behave different but have exactly the same Bayes error, which I am going to ignore). The second feature that is selected by feature forward selection is, in combination with that first feature, by construction, at least as good as the second feature from the individual selection. Maybe good to reason from contradiction?

(c) One can construct examples where individual is better than feature forward, but one needs at least 4 dimensions and one needs to reduce to 3D. Roughly, one can make a problem that is perfectly to solve in 3 of the 4 dimensions where these three dimensions individually rank as the top three. However, one can mislead feature forward selection and have it choose a second feature that at that point gives more improvement in terms of the Byes error, but that does not lead to separable classes when reaching 3 dimensions. A specific example can be obtained for two Gaussian classes, one with mean (0,0,0,0) and one with mean (1,1,1,1), in case one chooses the covariance matrices to be the same and equal to, for instance, (7 -3 -3 -4; -3 8 -3 4; -3 -3 8 4; -4 4 4 9).

Exercise 6.16
(a) $10 + 9 + 8 = 27$
(b) $10 + 9 + 8 + 7 + 6 + 5 + 4 = 49$?
(c) $\binom{10}{3} = 120$
(d) I think both $2^10 - 1 = 1023$ and $2^10 = 1024$ would do for me.
(e) Fisher is scaling independent and so need to check choosing 1, 2, and 3 features, which makes $\binom{10}{3} + \binom{10}{2} + \binom{10}{1} = 120 + 45 + 10 = 175$. For NMC the number of copies might matter, so... we could "enumerate" all possibilities? 1 2 3 ... 10 = 10; 12 13 14 15 ... 110 23 24 25 ... 210 34 ... 910 = $\binom{10}{2} = 45$ (times two because either first or last should be taken twice); and then finally just $\binom{10}{3} = 120$. So this makes $10 + 2 \cdot 45 + 120 = 220$?

Exercise 6.17
75,287,520

Exercise 6.21
(a) Feature 1.
(b) Obviously feature 1 again.
(c) between-class covariance is $\begin{pmatrix} 6\frac{1}{3} & 4 \\ 4 & 3 \end{pmatrix}$ (or some multiple of this matrix) and explicit multiplication checks that both vectors are indeed eigenvectors.
(d) Within-class is of 'no' influence and solution is based on eigenvalue decomposition of the between-class alone, which basically has been solved in the previous question.
(e) feature extraction performs worst and the feature selection attains an overlap of 0.
(f) E.g. all class means on one line not aligned with one of the axes.
(g) makes no difference whatsoever.

Exercise 6.22
(a) There are many possible solutions, though the "means on grid" restriction limits the possibilities drastically. One solution is given by $(3, 1)$, $(1, 3)$, and the third disc somewhere on $(5, 5)$, or $(6, 6)$, etc.
(b) No, this is impossible. For them to overlap perfectly, they have to be perfectly aligned, but this means that exactly the wrong, i.e., the most optimal direction, will be chosen by Fisher.
(c) There are many possible solutions, though the "means on grid" restriction limits the possibilities. The only thing one needs to do is pick any rectangle aligned with the axes and put the cluster means in its vertices making sure that clusters from one class are in opposing vertices.
(d) Yes.
(e) "Obviously" not.

Exercise 6.23
(a) "Inkomertje" : 0
(b) Feature A : $1^2 + 2^2 + 3^2 = 14$, Feature B : $0^2 + 6^2 + 6^2 = 72$, so we would choose Feature B, because the largest summed distance
(c) Feature A : $\frac{1}{6}$, Feature B : $\frac{1}{3}$, so we would choose Feature A, because the error is smaller
(d) Same answers as in b.
(e) Feature A : $\frac{5}{12}$, Feature B : $\frac{4}{12}$, so we would now choose Feature B instead of Feature A
(f) $\begin{pmatrix} 14/9 & -24/9 \\ -24/9 & 72/9 \end{pmatrix}$... but any answer the is proportional will do for me

(g) Add up $\begin{pmatrix} 14/9 & -24/9 \\ -24/9 & 72/9 \end{pmatrix}$ and $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$

Exercise 6.24
(a) bagging indeed helps in certain setting. E.g. try
cls = treec([],'infcrit',0)
prwaitbar off
T1 = 0; T2 = 0;
for i = 1:10
trn = gendats([50 50]);
tst = gendats([1000 1000]);
w1 = trn*cls;
w2 = baggingc(trn,cls);
T1 = T1 + tst*w1*testc;
T2 = T2 + tst*w2*testc;
end
T1/10,T2/10
(b) Rerun the previous code with the different data set.

Exercise 6.25
(a) Well, do I really need say more?
(b) Compute, for instance, `c*nmc(b)*testc`.
(c) The min combiner and, especially, the product combiner seem to offer some improvements. Max and median seem to make stuff worse. Mean doesn't do much.
(d) It is crucial that, especially in the test phase, we combine features coming from the same object. If we would have applied gendat separately to all data sets, there would be no guarantee that the same objects are in all training sets or test sets, respectively.

Exercise 6.26
(a) There are four features, coming from the two posteriors from the two classifiers.
(b) They correspond to the two posteriors of belonging to the first class from the two classifiers.
(c) Typically, better classification performance seems possible as we can find a linear classifier in the posterior space that performs better than both the horizontal and vertical lines. This result might, however, depend on the linear classifiers actually chosen.

Exercise 6.27
(a) The combiner typically performs slightly better. The reason for this is, however, somewhat disappointing as it doesn't seem to pick up any true nonlinearities in the decision boundary.

Exercise 6.28
(a) We see the decision boundary become progressively more complex.
(b) The decision boundary (once in a while) becomes overly complex and starts to fit to insignificant details in the training data. The test error would slowly go up again with more base classifiers after initially decreasing.