Martin Lundfall, Henri Bunting, Malte Siemers, Patrik Bey

Exercise sheet 07 - Machine Intelligence I

7.1

Given the common covariance matrix Σ of the two classes, we find the eigenvalues of the matrix through eigenvalue decomposition (find the solutions of λ of $det(A - \lambda I) = 0$) and set the weights to $w_i := \lambda_i$ for each dimension $i \in \{1, ..., d\}$. We then get b from the average of the conditional means of the two classes $b = (\mu_1 - \mu_2)/2$.
If we have two different covariance matrices for the two classes then the slope of the hyperplane won't simply be the eigenvalues of the covariance matrix, but rather by the difference of the means weighted by the sum of the covariances as given by Fisher:
$w \propto (\Sigma_0 + \Sigma_1)^{-1}(\mu_1 - \mu_0)$

7.3

a)

The binomial distribution can be a good model for the probability of a particulair sequence of binary outcomes, such as the probability of getting k heads in a n coinflips. In the case of a fair coin we have $p = \frac{1}{2}$. Its central properties are that p is constant for each trial k and that we only have two possible outcomes (hence binomial).

b)

Whenever $p = \frac{1}{2}$ the approximation is reasonable, but also as n becomes sufficiently large in relation the distance $|p - \frac{1}{2}|$. When p is close to either end of the spectrum $[0, 1]$, the distribution becomes skewed for small n. But we can see that higher order terms of $f(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$ tend to zero as n gets sufficiently large, thus making the distribution symmetric around its mean $k = np$.
One reason why this is so widely used is because it is repeatedly found in nature, and very suitable for random variables that are subject to noise. It is a good approximation for the binomial distribution of the coin flip example given above, since our coin is fair, $p = \frac{1}{2}$, for any n.

c)

The skewness of the Poisson Distribution (PD) is $\lambda^{-\frac{1}{2}}$. The skewness of the Binomial Distribution (BD) equals to $\frac{1-2p}{\sqrt{np(1-p)}}$. A BD with $p < 0.5$ and small n and a PD with small values of λ would have a similar form and a PD with large λ could approximate a BD with sufficiently large n. In both cases the means should be similar: $\lambda \approx np$. In the coin flip experiment, we have $p = 0.5$. That yields a skewness of 0 in the BD. So the PD would only give good predictions for lots of coin flips.