# CO902
# Solutions to Problem Set 4

1. *Estimating "a priori" probabilities.* Using the suggested representation $\{\mathbf{x}_i, \mathbf{t}_i\}$, the $\mathbf{t}_i$ are samples from a multinomial distribution with success probability parameter vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ and sample size of 1 (the counts $\sum_{i=1}^n \mathbf{t}_i$ are multinomial with sample size of $n$, but we'll stick with the individual $\mathbf{t}_i$'s).

   The joint likelihood of the predictor and responses is

$$
\begin{aligned}
p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{t}_1, \ldots, \mathbf{t}_n) &= p(\mathbf{x}, \mathbf{t}) && (1)\\
&= p(\mathbf{x}|\mathbf{t})p(\mathbf{t}) && (2)\\
&= \left(\prod_{i=1}^n \prod_{k=1}^K p_k(\mathbf{x}_i)^{t_{ik}}\right)\left(\prod_{i=1}^n \prod_{k=1}^K \pi_k^{t_{ik}}\right) && (3)
\end{aligned}
$$

   where $p_k(\mathbf{x}_i)$ is the class $k$ conditional distribution. To find the maximum likelihood estimator of $\boldsymbol{\pi}$, we must account for the constraint that $\sum_{k=1}^K \pi_k = 1$, and so consider optimizing log joint likelihood plus a Lagrangian term

$$
\sum_{i=1}^n \sum_{k=1}^K t_{ik} \log(p_k(\mathbf{x}_i)) + \sum_{i=1}^n \sum_{k=1}^K t_{ik} \log(\pi_k) + \lambda(\sum_{k=1}^K \pi_k - 1), \quad (4)
$$

   and then taking derivative w.r.t. $\pi_k$ and $\lambda$, equating to zero and solving...

$$
\begin{aligned}
0 = \frac{\partial}{\partial \pi_k} \log p(\mathbf{x}, \mathbf{t}) &= 0 + \sum_{i=1}^n t_{ik}/\pi_k + \lambda, && (5)\\
\Rightarrow \hat{\pi}_k &= -\frac{1}{\lambda} \sum_{i=1}^n t_{ik}; && (6)
\end{aligned}
$$

   now, note that the Langrangian gives us

$$
\begin{aligned}
1 = \sum_{k=1}^K \hat{\pi}_k &= -\frac{1}{\lambda} \sum_{k=1}^K \sum_{i=1}^n t_{ik} && (7)\\
&= -\frac{1}{\lambda} n, && (8)\\
&\Rightarrow \hat{\lambda} = -n, && (9)
\end{aligned}
$$

   because $\sum_{k=1}^K t_{ik} = 1$, and this gives the final result that $\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^n t_{ik}$.

   In summary, the key result here is that the estimated incidence of class $k$ is simply the proportion of samples in that class, and this estimate is the same regardless of the class conditional distribution.

2. *Optimal decision rule for continous data.* Follows identically to the class notes "Decision Theory Background of Classification", but with continuous random variables.

   We seek optimal decision rule $D_{\mathbf{x}}$ maximizes the probability

$$\begin{aligned} P(D_{\mathbf{x}} = Y) &= \sum_{k=1}^{K} P(D_{\mathbf{x}} = k, Y = k) \\ &= \sum_{k=1}^{K} P(\mathbf{x} \in \mathcal{R}_k, Y = k) \\ &= \sum_{k=1}^{K} \int_{\mathbf{x} \in R_k} P(\mathbf{x}, Y = k) d\mathbf{x} \\ &= \sum_{k=1}^{K} \int_{\mathbf{x} \in R_k} P(Y = k|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

   where $R_k \subset \Re^d$ are the values of $x$ for which decision $k$ is to be made. The challenge, then, is to construct $R_k$ by choosing, for each $x$, which $R_k$ it should belong to so as to maximize this expression.

   The key here is to see that, for a given $x$, $p(\mathbf{x})$ is the same for all $k$, and so to maximise this expression we need to see $P(Y = k|\mathbf{x})$ as a function of $x$, and we need to choose the $k$ that maximises $P(Y = k|\mathbf{x})$. Thus

$$R_k = \{\mathbf{x} : P(Y = k|\mathbf{x}) \geq P(Y = k'|\mathbf{x}) \text{ for } k' \neq k\},$$

   or, equilvalently the optimal decision rule is

$$D_{\mathbf{x}} = \hat{Y}(\mathbf{x}) = \arg\max_k P(Y = k|\mathbf{x}).$$

   Of course, in practice we usually rely on Bayes Rule to express $P(Y = k|\mathbf{x})$ in terms of the class-conditional distributions $p(\mathbf{x}|Y = k)$, and of course (of course) these distributions must be estimated in order to make a real live working classifier.

3. *Regression.*

   If you avail yourself of matrix-mode calculations, these are very consise results.

   (a) Algebraic derivation of OLS estimates.

$$\begin{aligned} 0 = \frac{\partial}{\partial \mathbf{w}} J(\mathbf{w}) &= \frac{\partial}{\partial \mathbf{w}} \left( \mathbf{Y}^\top \mathbf{Y} - 2(\mathbf{X}\mathbf{w})^\top \mathbf{Y} + (\mathbf{X}\mathbf{w})^\top (\mathbf{X}\mathbf{w}) \right) && (10) \\ &= 0 - 2\frac{\partial}{\partial \mathbf{w}} \mathbf{w}^\top \mathbf{X}^\top \mathbf{Y} + \frac{\partial}{\partial \mathbf{w}} \mathbf{w}^\top (\mathbf{X}^\top \mathbf{X}) \mathbf{w} && (11) \\ &= -2\mathbf{X}^\top \mathbf{Y} + 2\mathbf{X}^\top \mathbf{X}\mathbf{w} && (12) \\ \Rightarrow \quad \mathbf{X}^\top \mathbf{Y} &= \mathbf{X}^\top \mathbf{X}\mathbf{w} && (13) \end{aligned}$$

The last expression represents the Normal Equations, and any $\mathbf{w}$ that satisfies this expression is a minimizer of the residual sum of squares. Assuming $\mathbf{X}$ is of full rank, the solution is $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$; if $\mathbf{X}$ is not of full rank, the Moore-Penrose pseudo inverse of $\mathbf{X}$ can produce one of the infinite number of solutions, $\hat{\mathbf{w}} = \mathbf{X}^-\mathbf{Y}$.

(b) Maximum Likelihood derivation of OLS estimates based on Normality.
The likelihood of the data is

$$p(\mathbf{Y}|\mathbf{w},\sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{1}{2}(\mathbf{Y}-\mathbf{X}\mathbf{w})^\top(\mathbf{Y}-\mathbf{X}\mathbf{w})/\sigma^2\right)$$

and the log likelihood, dropping additive constants that don't depend on $\mathbf{w}$, is

$$-\frac{1}{2}(\mathbf{Y}-\mathbf{X}\mathbf{w})^\top(\mathbf{Y}-\mathbf{X}\mathbf{w})/\sigma^2 = -\frac{1}{2}J(\mathbf{w})/\sigma^2$$

i.e. $J(\mathbf{w})$ from the previous part; thus the value of $\mathbf{w}$ that minimized $J(\mathbf{w})$ will maximize the log liklihood, and hence the MLE is again $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ for full-rank $\mathbf{X}$ and $\hat{\mathbf{w}} = \mathbf{X}^-\mathbf{Y}$ otherwise.

4. *Ridge Regression.* For clarity, refer to the Ridge objective function as $J_R(\mathbf{w})$.

(a) Algebraic derivation of Ridge Regression estimator

$$\begin{aligned}
0 = \frac{\partial}{\partial\mathbf{w}}J_R(\mathbf{w}) &= \frac{\partial}{\partial\mathbf{w}}\left(\mathbf{Y}^\top\mathbf{Y} - 2(\mathbf{X}\mathbf{w})^\top\mathbf{Y} + (\mathbf{X}\mathbf{w})^\top(\mathbf{X}\mathbf{w}) + \lambda\mathbf{w}^\top\mathbf{w}\right) && (14)\\
&= 0 - 2\frac{\partial}{\partial\mathbf{w}}\mathbf{w}^\top\mathbf{X}^\top\mathbf{Y} + \frac{\partial}{\partial\mathbf{w}}\mathbf{w}^\top(\mathbf{X}^\top\mathbf{X})\mathbf{w} + \lambda\frac{\partial}{\partial\mathbf{w}}\mathbf{w}^\top\mathbf{w} && (15)\\
&= -2\mathbf{X}^\top\mathbf{Y} + 2\mathbf{X}^\top\mathbf{X}\mathbf{w} + 2\lambda\mathbf{w} && (16)\\
\Rightarrow \quad & \mathbf{X}^\top\mathbf{Y} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\mathbf{w} && (17)
\end{aligned}$$

Again, if $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ is invertable (and for large enough $\lambda$, it always will be), then $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}$. In this case, the "trick" of using the Moore Penrose inverse directly with $\mathbf{X}$ doesn't work, but should $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ be rank-deficient then one could compute $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^-\mathbf{X}^\top\mathbf{Y}$; however, but it would defeat the purpose of Ridge Regression, which is increasing $\lambda$ until stable estimates are obtained when the pseudoinverse won't be needed.

(b) Maximum Likelihood derivation of OLS estimates based on Normality.
The posterior of $\mathbf{w}$ given the data $\mathbf{Y}$ is (regarding $\sigma^2$ as fixed)

$$\begin{aligned}
p(\mathbf{w}|\mathbf{Y}) = & \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(-\frac{1}{2}(\mathbf{Y}-\mathbf{X}\mathbf{w})^\top(\mathbf{Y}-\mathbf{X}\mathbf{w})/\sigma^2\right)\times && (18)\\
& \frac{1}{(2\pi)^{1/2}\sigma_0}\exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{0})^\top(\mathbf{w}-\mathbf{0})/\sigma_0^2\right) && (19)
\end{aligned}$$

and the log posterior, dropping additive constants that don't depend on $\mathbf{w}$, is

$$
\begin{aligned}
&-\frac{1}{2}(\mathbf{Y}-\mathbf{X}\mathbf{w})^\top(\mathbf{Y}-\mathbf{X}\mathbf{w})/\sigma^2+\frac{1}{2}\mathbf{w}^\top\mathbf{w}/\sigma_0^2 \\
&\qquad\qquad\propto(\mathbf{Y}-\mathbf{X}\mathbf{w})^\top(\mathbf{Y}-\mathbf{X}\mathbf{w})+\frac{\sigma^2}{\sigma_0^2}\mathbf{w}^\top\mathbf{w}.
\end{aligned}
\tag{20}
$$

This is of course $J_R(\mathbf{w})$, with $\lambda=\sigma^2/\sigma_0^2$. The interpretation is that when our prior belief on $\mathbf{w}$ is vague relative to the residual noise, i.e. $\sigma_0 \gg \sigma$, then we need to trust the data, the $\mathbf{w}^\top\mathbf{w}$ penality is diminshed and we should get estimates close to OLS; if we have strong prior belief that $\mathbf{w}$ should close to zero, then $\sigma_0 \ll \sigma$, $\lambda$ will be large, then $\mathbf{w}^\top\mathbf{w}$ penality is active and $\mathbf{w}$ estimates will be shunk towards 0 relative to OLS's estimates.