1 Uniform convergence and Model Selection

a) The hypotheses $\hat{h}_i$ form a new hypothesis class of size k, with generalization error estimated on a dataset of size βm. We can therefore plug $\frac{\delta}{2}$ in for δ the formula in Part 6, page 7, and βm in for m, to get

$$|\hat{\varepsilon}(h_i) - \hat{\varepsilon}_{cv}(h_i)| \leq \sqrt{\frac{1}{2\beta m} \log\left(\frac{4k}{\delta}\right)}.$$

b) This result is analogous to the Theorem from Part 6, page 7, with $m = \beta m$, $\frac{\delta}{2}$ in place of δ, and with the coefficient 2 pulled underneath the square root.

c) This result follows immediately by solving for $\varepsilon\left(\hat{h}_j\right)$ and plugging the solution in for $\min_{i=1,\ldots k} \varepsilon\left(\hat{h}_i\right)$ in the equation from part b.

2 VC Dimension

a) This hypothesis class can shatter a single point x by letting $a < x$ if $x = 0$ and $a > x$ otherwise. It cannot shatter two points with different signs, so its VC dimension is 1.

b) This hypothesis class can shatter two points x,y. Assume $x < y$. If $f(x) = f(y) = 1$, then let $a < x < y < b$. If $f(x) = f(y) = 0$, then let $x < a < b < y$. If $f(x) = 0$ and $f(y) = 1$, then let $x < a < y < b$, and if $f(x) = 1$ and $f(y) = 0$, then let $a < x < b < y$. It cannot shatter three points x, y, z,where $x < y < z$, and $f(x) = f(z) = 1$, and $f(y) = 0$.

c) This hypothesis class can shatter a single point by alternating the sign of a. With two points, however, the hypothesis class can shatter pairs that are the same sign or different signs, but there is not a hypothesis that can shatter both. The VC dimension is therefore 1.

d) Given the periodicity of the sin function, we only have to consider the hypothesis class between 0 and 2π. Along this interval, the class is equivalent to the hypothesis class from part b, so it's VC dimension is 2.

3 l_1 regularization for least squares

a) Let $\theta_{\overline{k}}$ indicate the vector θ with the kth element set to zero, and X_k indicate the kth column of the matrix X. Then we can write the objective function $J(\theta)$ as

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}\left(\theta_{\bar{k}}x^{(i)} + \theta_k x_k^{(i)} - y^{(i)}\right)^2 - \lambda\sum_{i=1}^{n}|\theta_i|$$

Taking the derivative with respect to θ_k, we get

$$\begin{aligned}\frac{\partial J}{\theta_k} &= \sum_{i=1}^{m} \left(\theta_{\bar{k}} x^{(i)} + \theta_k x_k^{(i)} - y^{(i)}\right) x_k^{(i)} + s\lambda \\ &= X_k^T \left(X\theta_{\overline{k}} + X_k\theta_k - Y\right) + s\lambda\end{aligned}$$

Solving for θ_k, we get

$$\theta_k = \left(X_k^T X_k\right)^{-1} \left[X_k^T Y - s\lambda - X_k^T X_k \theta_k\right]$$

To implement coordinate descent with ℓ_1 regularization, we would calculate θ_k for s $= \pm 1$, plug each value of θ_k back into $J(\theta)$, and choose the value that maximizes the objective function.

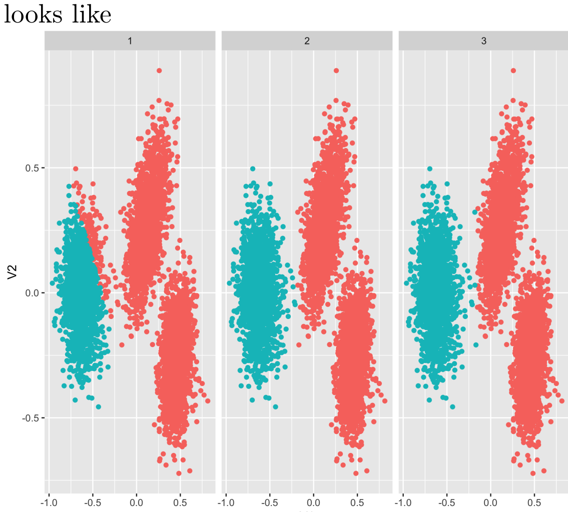
b) An implementation of the above algorithm is in the q3.R file.

c) ℓ_1 regularization results in sparse parameter vectors θ. The features to select are precisely those features that correspond to the indices with non-zero values in θ.

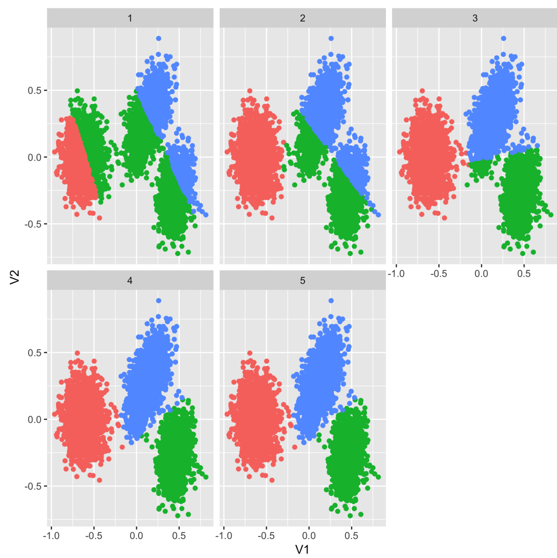
4 K-Means Clustering

An implementation of k-means is located in the file q4.R.

k=2 looks like



k=3 looks like



5 The generalized EM algorithm

a) The argument for convergence is nearly identical to the one given in lecture for the EM algorithm. Write the objective function as $J(Q, \theta)$. Then $\theta_{t+1} := \theta_t + \alpha \nabla_\theta J$, where the learning rate, α, is chosen small enough to ensure that $J(Q, \theta_{t+1}) \geq J(Q, \theta_t)$. Then we have that

$$\begin{aligned} \ell(\theta_{t+1}) &\geq J(Q, \theta_{t+1}) \\ &\geq J(Q, \theta_t) \\ &= \ell(\theta_t), \end{aligned}$$

so the likelihood function is monotonically increasing. The only difference between this derivation and the derivation in the notes is that the second inequality is now justified by the choice of α.

b) Calculating the partial derivative $\nabla_\theta \sum_i \log \sum_{z^{(i)}} p\left(x^{(i)}, z^{(i)}; \theta\right)$, we get

$$\begin{aligned}
\sum_i \frac{1}{\sum_{z^{(i)}} p\left(x^{(i)}, z^{(i)}; \theta\right)} \nabla_\theta \sum_{z^{(i)}} \frac{Q_i}{Q_i} p\left(x^{(i)}, z^{(i)}; \theta\right) \\
&= \sum_i \sum_{z^{(i)}} \frac{Q_i}{Q_i} \frac{1}{p\left(x^{(i)}, z^{(i)}; \theta\right)} \nabla_\theta p\left(x^{(i)}, z^{(i)}; \theta\right) \\
&= \sum_i \sum_{z^{(i)}} \frac{Q_i}{Q_i} \nabla_\theta \log p\left(x^{(i)}, z^{(i)}; \theta\right) \\
&= \nabla_\theta \sum_i \sum_{z^{(i)}} Q_i \log \frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i}.
\end{aligned}$$