2. Regularized and Kernel k-means.

(a). 0. we just put each point as its own cluster.

(b). Let $f(\mu_i) = \lambda \|\mu_i\|_2^2 + \sum_{x_j \in C_i} \|x_j - \mu_i\|_2^2$.

$$\nabla_{\mu_i} f(\mu_i) = 2\lambda \mu_i + \left(2 \sum_{x_j \in C_i} (\mu_i - x_i)\right)$$

$$= 2\left((|C_i| + \lambda)\mu_i - \sum_{x_j \in C_i} x_j\right) = 0$$

we have $\hat{\mu}_i = \frac{1}{|C_i| + \lambda} \sum_{x_j \in C_i} x_j$

the optimum is obtained at $\hat{\mu}_i$ since f is a convex function.

(c).
$$\min_{\mu_i \in \mathbb{R}^d} \sum_{i=1}^{k} \left(\|\mu_i\|_2 + \sum_{x_j \in C_i} \|x_j - \mu_i\|_2\right).$$

(d) For a cluster S_i, we let $\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} \Phi(x)$.

and we want to minimize $\sum_{x \in S_i} \|\Phi(x) - \mu_i\|_2^2$. for $i \in 1 \cdots K$.

for k th cluster, we have

$$f(i,k) = \|\Phi(x_i) - \mu_k\|^2 = \Phi(x_i)\cdot\Phi(x_i) - 2\Phi(x_i)\cdot\mu_k + \mu_k\cdot\mu_k$$

$$= \Phi(x_i)\Phi(x_i) - \frac{2}{|S_k|}\sum_{x_j \in S_k} \Phi(x_i)\Phi(x_j) + \frac{1}{|S_k|^2}\sum_{x_j \in S_k}\sum_{x_m \in S_k} \Phi(x_j)\Phi(x_m)$$

$$= K(x_i, x_i) - \frac{2}{|S_k|}\sum_{x_j \in S_k} \Phi(x_i)\cdot\Phi(x_j) + \frac{1}{|S_k|^2}\sum_{x_j \in S_k}\sum_{x_m \in S_k} K(x_j, x_m)$$

Hence

$$\text{class}(i) = \arg\min_k \frac{1}{|S_k|^2}\sum_{x_j, x_m \in S_k} K(x_j, x_m) - \frac{2}{|S_k|}\sum_{x_j \in S_k} K(x_i, x_j)$$

$$+ K(x_i, x_i)$$

(e).

1. We can drop $K(x_i, x_i)$ as it's not related to k.

2. For the symmetric kernels, we need only compute it for once

3. For all the kernel, we need only compute them at the beginning of the algorithm, store them (probabally in a numpy matrix) hence we don't need to compute them again within the looping.

Scanned with CamScanner

3. The Training Error of AdaBoost.

(a). $$W_i^{(T+1)} = \frac{W_i^{(T)} \exp(-\beta_T y_i G_T(X_i))}{Z_T}$$

$$\sum_{i=1}^{n} W_i^{(T+1)} = \sum_{i=1}^{n} \frac{W_i^{(T)} \exp(-\beta_T y_i G(X_i))}{Z_T}$$

$$1 = \frac{1}{Z_T} \left(\sum_{y_i = G(x_i)} W_i^{(T)} e^{-\beta_T} + \sum_{y_i \neq G(x_i)} W_i^{(T)} e^{\beta_T} \right)$$

$$Z_T = \sum_{y_i = G(x_i)} W_i^{(T)} e^{-\beta_T} + \sum_{y_i \neq G(x_i)} W_i^{(T)} e^{\beta_T}.$$

$$= \sum_{y_i = G(x_i)} W_i^{(T)} \left(\frac{1-err_T}{err_T}\right)^{-\frac{1}{2}} + \sum_{y_i \neq G(x_i)} W_i^{(T)} \left(\frac{1-err_T}{err_T}\right)^{\frac{1}{2}}$$

$$= \left(\frac{1-err_T}{err_T}\right)^{\frac{1}{2}} \left((1-err_T) \frac{err_T}{1-err_T} + err_T \right)$$

$$= 2\sqrt{err_T(1-err_T)}$$

(b) Let $M(X_i) = \sum_{t=1}^{T} \beta_t G_t(X_i)$.

$$W_i^{(T+1)} = \frac{1}{Z_T} W_i^{(T)} e^{-\beta_T y_i G_T(X_i)} = \frac{1}{Z_T \cdot Z_{T-1}} W_i^{(T-1)} e^{-y_i (\beta_T G_T(X_i) + \beta_{T-1} G_{T-1}(X_i))}$$

$$= \frac{1}{n} \cdot \frac{1}{\prod_{t=1}^{T} Z_t} \cdot e^{-y_i M(X_i)}.$$

(c) $$\sum_{i=1}^{n} e^{-y_i M(X_i)} = \sum_{\substack{sign(M(X_i)) \\ = y_i}} e^{-y_i M(X_i)} + \sum_{\substack{sign(M(X_i)) \\ \neq y_i}} e^{-y_i M(X_i)}.$$

$$\geq \sum_{\substack{sign(M(X_i)) \\ \neq y_i}} e^{-y_i M(X_i)} \geq \sum_{\substack{sign(M(X_i)) \\ \neq y_i}} 1 = B.$$

we have the last inequality because for any $x \geq 0$, $e^x \geq 1$.

Scanned with CamScanner

(d) $err_T \leq 0.49 \Rightarrow Z = 2\sqrt{err_T(1-err_T)} \leq 0.9998$.

From (5), we have $\sum_{i=1}^{n} W_i^{(T+1)} = \sum_{i=1}^{n} \frac{1}{n\prod_{t=1}^{n} Z_t} e^{-y_i M(X_i)}$

$$\geq \sum_{i=1}^{n} \frac{1}{n(0.9998)^T} e^{-y_i M(X_i)}$$

From (6)

$$n(0.9998)^T \cdot \sum_{i=1}^{n} W_i^{(T+1)} \geq \sum_{i=1}^{n} e^{-y_i M(X_i)} \geq B.$$

$\lim_{T\to 0} n(0.9998)^T \sum_{i=1}^{n} W_i^{(T+1)} = 0$, but $B \geq 0$.

we have $\lim_{T\to 0} B = 0$.

(e) Eeach short decision tree only pick several features, combining Adaboost, those trees with "bad features" (features that don't perfrom well) will have a small weight and essentially don't get used in the final decision.

Hence, this is the same as subset selecting features.