# Machine Learning

Fabio Vandin

January $17^{th}$, 2023

Exam example 1, exercise 4, part b

Proof. We consider the cost function of the k-means problem as a function of the centroids $\vec{\mu}_1, \dots, \vec{\mu}_k$: $f(\vec{\mu}_1, \vec{\mu}_2, \dots, \vec{\mu}_k) = \sum_{i=1}^{k} \sum_{\vec{x} \in C_i} d(\vec{x}, \vec{\mu}_i)^2$

At the optimum, the gradient is equal to $\vec{0}$ (and it is a convex function, so the global optimum is the only local optimum).

Let's compute the gradient: $\left[ \frac{\partial f}{\partial \vec{\mu}_1}, \frac{\partial f}{\partial \vec{\mu}_2}, \dots, \frac{\partial f}{\partial \vec{\mu}_k} \right]$.

Let's consider $\frac{\partial f}{\partial \mu_j}$:

$$\frac{\partial \left( \sum_{i=1}^{k} \sum_{\vec{x} \in C_i} d(\vec{x}, \vec{\mu}_i)^2 \right)}{\partial \vec{\mu}_j} = \sum_{i=1}^{k} \frac{\partial \left( \sum_{\vec{x} \in C_i} d(\vec{x}, \vec{\mu}_i) \right)}{\partial \vec{\mu}_j} =$$

$$= \frac{\partial \left( \sum_{\vec{x} \in C_j} d(\vec{x}, \vec{\mu}_j)^2 \right)}{\partial \vec{\mu}_j} = \sum_{\vec{x} \in C_j} \frac{\partial \left( d(\vec{x}, \mu_j)^2 \right)}{\partial \vec{\mu}_j} =$$

$$= \sum_{\vec{x} \in C_j} \frac{\partial \left( (\vec{x} - \vec{\mu}_j)^T (\vec{x} - \vec{\mu}_j) \right)}{\partial \vec{\mu}_j} =$$

$$= \sum_{\vec{x} \in C_j} \left( +2\vec{\mu}_j - 2\vec{x} \right) = \left( \sum_{\vec{x} \in C_j} 2\vec{\mu}_j \right) - \left( \sum_{\vec{x} \in C_j} 2\vec{x} \right)$$

$$= 2 \, |C_j| \, \vec{\mu}_j - 2 \sum_{\vec{x} \in C_j} \vec{x}$$

At the optimum:

$$2|C_j|\vec{\mu}_j - 2\sum_{\vec{x}\in C_j}\vec{x} = \vec{0}$$

$$\iff \cancel{2}|C_j|\vec{\mu}_j = \cancel{2}\sum_{x\in C_j}\vec{x}$$

$$\iff \vec{\mu}_j = \frac{1}{|C_j|}\sum_{x\in C_j}\vec{x}$$

(that is the update rule for $\mu_j$ in Lloyd's alg.) $\square$

Exam example 2, exercise 2, part b

For SGD we need to compute $\nabla \ell(h(x),(x,y))$.
Since $\ell()$ (the loss function) is defined by 2 cases
$(y=1$ and $y=0)$, the $\nabla \ell()$ is defined by the
same 2 cases $(y=1$ and $y=0)$.
We consider the two cases i) $y=1$ and ii) $y=0$

i) $y=1 \Longrightarrow \nabla \ell(h(x),(x,y)) = \left[\frac{\partial \ell}{\partial w_1}, \frac{\partial \ell}{\partial w_2}\right]^T$. Let's
compute $\frac{\partial \ell}{\partial w_1}$ and $\frac{\partial \ell}{\partial w_2}$:

$$\frac{\partial \ell}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \ell}{\partial z} = \frac{\partial (w_1 + w_2 x^2)}{\partial w_1} \cdot \frac{\partial \left(1 - \frac{1}{1+e^{-z}}\right)}{\partial z}$$

see the text of the exercise $(z = w_1 + w_2 x^2)$

$$= 1 \cdot \left( -\frac{e^z}{(1+e^z)^2} \right)$$

$$= -\frac{e^z}{(1+e^z)^2}, \text{ with } z = w_1 + w_2 x^2$$

$$\frac{\partial \ell}{\partial w_2} = \frac{\partial z}{\partial w_2} \cdot \frac{\partial \ell}{\partial z}$$

$$= x^2 \left( -\frac{e^z}{(1+e^z)^{2}} \right)$$

$$= -x^2 \frac{e^z}{(1+e^z)^2}$$

$$d\left(\frac{1}{1+e^{-z}}\right) = ?$$

$$d\left(\frac{f}{g}\right) = \frac{(df)g - f(dg)}{g^2}$$

$$\frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z}$$

$$\Rightarrow \frac{d}{dz}\left(\frac{e^z}{1+e^z}\right)$$

$$= \frac{e^z(1+e^z) - e^z \cdot e^z}{(1+e^z)^2} =$$

$$= \frac{e^z + e^{2z} - e^{2z}}{(1+e^z)^2} = \frac{e^z}{(1+e^z)^2}$$

ii) $y = 0 \Rightarrow \nabla \ell(h(x), (x,y)) = \left[\frac{\partial \ell}{\partial w_1}, \frac{\partial \ell}{\partial w_2}\right]^T.$

$$\frac{\partial \ell}{\partial w_1} = \frac{\partial z}{\partial w_1} \cdot \frac{\partial \ell}{\partial z} = 1 \cdot \frac{\partial \left( \frac{1}{1+e^{-z}} \right)}{\partial z} = \frac{e^{z}}{(1+e^{z})^2}$$

$$\frac{\partial \ell}{\partial w_2} = \frac{\partial z}{\partial w_2} \cdot \frac{\partial \ell}{\partial z} = x^2 \cdot \frac{e^{z}}{(1+e^{z})^2}$$

Therefore the SGD update rule is:

- pick $(x_i, y_i) \in S$ uniformly at random
- if $y_i = 1$ then $\vec{w}^{(t+1)} \leftarrow \vec{w}^{(t)} + \eta \begin{bmatrix} e^{z}/(1+e^{z})^2 \\ x^2\, e^{z}/(1+e^{z})^2 \end{bmatrix}$;

  else $\vec{w}^{(t+1)} \leftarrow \vec{w}^{(t)} - \eta \begin{bmatrix} e^{z}/(1+e^{z})^2 \\ x^2\, e^{z}/(1+e^{z})^2 \end{bmatrix}$;