Example Proofs

1 Representatives of points

1.1 L_2-distance optimum

Consider a set of d-dimensional points $X = \{x_1, \ldots, x_n\}$ and distance function

$$D_2(x_i, x_j) = \sum_{\ell=1}^{d} (x_i(\ell) - x_j(\ell))^2 .$$

Show that the representative

$$x^* = \arg\min_{x \in \mathbb{R}^d} \sum_{x_i \in X} D_2(x_i, x)$$

is the mean of the points in X. That is, for every $\ell \in \{1, \ldots, d\}$ $x^*(\ell) = \frac{1}{n} \sum_{i=1}^{n} x_i(\ell)$.

Answer We can differentiate this function and find its optimum. Since this function is convex up, it only has a minimum:

$$\begin{aligned}
\sum_{i=1}^{n} D_2(x_i, y) &= \sum_{i=1}^{n} \sum_{l=1}^{d} (x_{il} - y_l)^2 && \text{assumption} \\
\frac{\partial D_2}{y_j} &= \frac{\partial}{\partial y_j} \sum_{i=1}^{n} \sum_{l=1}^{d} (x_{il} - y_l)^2 && \text{differentiating both sides} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial y_j} \sum_{l} (x_{il} - y_l)^2 && \text{distribute derivative operator} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial y_j} (x_{ij} - y_j)^2 && \text{when } l \neq j \text{ derivative is } 0 \\
0 &= \sum_{i=1}^{n} 2(y_j - x_{ij}) && \text{derivative and setting derivative to } 0 \\
\sum_{i=1}^{n} x_{ij} &= \sum_{i=1}^{n} y_j && \text{algebra} \\
\sum_{i=1}^{n} x_{ij} &= n y_j && \sum_{i=1}^{n} 1 = n \\
\frac{1}{n} \sum_{i=1}^{n} x_{ij} &= y_j && \text{solving}
\end{aligned}$$

Since each element of y is independent from each other, this result stands for every element of y. Therefore, the optimal representative is the vector mean of all the points.

1.2 L_1-distance optimum

Consider a set of d-dimensional points $X = \{x_1, \ldots, x_n\}$ and distance function

$$D_1(x_i, x_j) = \sum_{\ell=1}^{d} |x_i(\ell) - x_j(\ell)| \,.$$

Show that the representative

$$x^* = \arg\min_{x \in \mathbb{R}^d} \sum_{x_i \in X} D_1(x_i, x)$$

is the median of the points in X. That is, for every $\ell \in \{1, \ldots, d\}$ $x^*(\ell) = \texttt{median}(x_1(\ell), \ldots, x_n(\ell)\}$.

Answer We can also solve this by differentiating. However, let us first break the data X into two sets. Let X^+ be the points in X that are above point y and let X^- be the points below point y:

$$\begin{aligned}
\sum_{i=1}^{n} D_1(x_i, y) &= \sum_{i=1}^{n} \sum_{l=1}^{d} |x_{il} - y_l| && \text{assumption} \\
\frac{\partial D_1}{y_j} &= \frac{\partial}{\partial y_j} \sum_{i=1}^{n} \sum_{l=1}^{d} |x_{il} - y_l| && \text{differentiating both sides} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial y_j} \sum_{l} |x_{il} - y_l| && \text{distribute derivative operator} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial y_j} |x_{ij} - y_j| && \text{when } l \neq j \text{ derivative is 0} \\
&= \sum_{x \in X^+} \frac{\partial}{\partial y_j} (x_j - y_j) + \sum_{x \in X^-} \frac{\partial}{\partial y_j} (-1)(x_j - y_j) && \text{splitting up points above and below } y_j \\
0 &= \sum_{x \in X^+} (-1) + \sum_{x \in X^-} 1 && \text{derivative and setting derivative to 0} \\
|X^+| &= |X^-| && \text{algebra}
\end{aligned}$$

Therefore the optimum occurs when half of the points are above and the other are below. This occurs when y_j is the median.

2 Locality Sensitive Hashing

Consider a locality sensitive hashing function h associated with distance function d. Assume that h and d are associated with the following relationship:

$$\Pr(h(x) \neq h(y)) = d(x, y)$$

for every pair of points x, y. Show that if the above equation is correct, then d is a metric.

Answer We need to show all 4 properties:
1) $Pr[\cdot] \in [0, 1] \rightarrow d(x, y) \geq 0$

2) We need to show both arrows:
$d(x, y) = 0 \rightarrow x = y$:

$$\begin{aligned} d(x,y) &= 0 && \text{assumption} \\ Pr[h(x) \neq h(y)] &= 0 && \text{def of } d \\ Pr[h(x) = h(y) &= 1 && Pr[h(x) = h(y)] = 1 - Pr[h(x) \neq h(y)] \\ x &= y && \text{collision will always occur only on equal items} \end{aligned}$$

$x = y \rightarrow d(x, y) = 0$:

$$\begin{aligned} x &= y && \text{assumption} \\ h(x) &= h(y) && x = y \\ Pr[h(x) = h(y)] &= 1 && \text{collision always occurs} \\ Pr[h(y) \neq h(x)] &= 0 && Pr[h(y) \neq h(x)] = 1 - Pr[h(x) = h(y)] \\ d(x,y) &= 0 && \text{plugging in} \end{aligned}$$

3) $d(x, y) = Pr[h(y) \neq h(x)] = Pr[h(x) \neq h(y)] = d(y, x)$

4) There are a few ways to show this. My favorite way is to use indicator random variables (IRVs). An indicator random variable takes on values 0 or 1. In general, IRVs are extremely useful. We will use it as such:
Let $I_{xy} = \begin{cases} 1 & Pr[h(x) \neq h(y)] \\ 0 & \text{otherwise} \end{cases}$
We can now write the triangle inequality as follows:

$$I_{xy} \leq I_{xz} + I_{yz}$$

We can prove this is true with a proof by contradition: Assume $I_{xy} > I_{xz} + I_{yz}$:

$$\begin{aligned} I_{xy} &\leq I_{xz} + I_{yz} && \text{assumption} \\ I_{xy} = 1 \cap I_{xz} &= 0 \cap I_{yz} = 0 && \text{only situation possible} \\ h(z) = h(z) \cap &\rightarrow h(x) = h(y) && I_{xz} = 0 \cap I_{xz} = 0 \rightarrow I_{xy} = 0 \end{aligned}$$

This is a contradiction because we know $I_{xy} = 1$. Therefore, this situation cannot occur.
Now that we know $I_{xy} \leq I_{xz} + I_{yz}$ is true, we can take the expected value of it:

$$\begin{aligned}
I_{xy} &\leq I_{xz} + I_{yz} && \text{assumption} \\
\mathbb{E}\Big[I_{xy} \leq I_{xz} + I_{yz}\Big] &= \mathbb{E}[I_{xy}] \leq \mathbb{E}[I_{xz}] + \mathbb{E}[I_{yz}] && \mathbb{E} \text{ is linear} \\
&= Pr[h(x) \neq h(y)] \leq Pr[h(x) \neq h(z)] + Pr[h(y) \neq h(z)] && \mathbb{E}[I] = Pr[I = 1] \\
&= d(x, y) \leq d(x, z) + d(y, z) && \text{abstraction}
\end{aligned}$$

Therefore, d is a metric