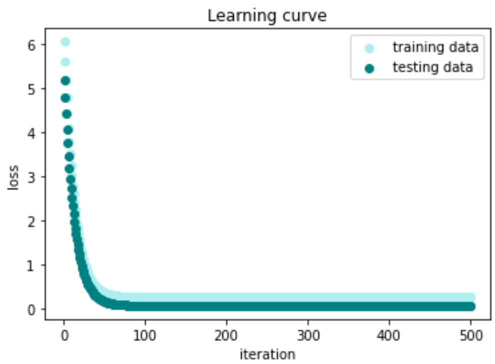
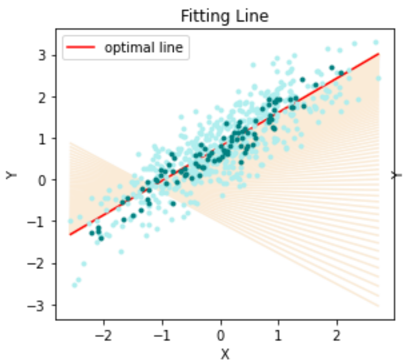
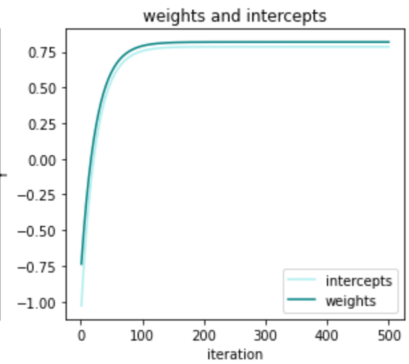
1. Learning curve of the training :



2. MSE = 0.06870297339333518
3. Fitting Line : Y = 0.8179703746689468X + 0.7845650803186925 (weight ~= 0.818, intercept ~= 0.785)





4. What's the difference between Gradient Descent,(GD) Mini-Batch Gradient Descent(Mini-BGD), and Stochastic Gradient Descent(SGD)?

	GD	Mini-BGD	SGD
Method	Parameters are updated after computing the gradient of error with respect to the entire dataset.	Parameters are updated after computing the gradient of error with respect to a subset of the dataset.	Parameters are updated after computing the gradient of error with respect to only a single data.
Noise	make smooth updates	Make less noisy updates than SGD (depend on the batch size)	make very noisy updates
Time	Take the most time	Balance between GD and SGD in terms of efficiency	Converges quickly for huge datasets.
convergence ratio	quick convergence ratio to a global minimum if the loss function is convex (and to local minimum one for non-convex functions)	Worse than GD	The worst (needs more iterations)
tolerance		Can "bounce around" global optimum — may require the bigger tolerance.	Can "bounce around" global optimum — may require the biggest tolerance.

1.

(1) $0.2\left(\frac{3}{10}\right) + 0.4\left(\frac{2}{4}\right) + 0.4\left(\frac{4}{20}\right) = 0.06 + 0.2 + 0.08$

$= 0.34$ ■

(2) $P(🍎 \wedge B \mid 🍎)$

$$= \frac{0.4(0.5)}{0.2(0.3) + 0.4(0.5) + 0.4(0.6)} = \frac{0.2}{0.5}$$

$= \frac{2}{5}$ ■

2. Let R_1 be the distribution area of class C_1, and R_2 be the area of class C_2, we know

① $P(\text{mistake}) = \int_{R_1} P(x, c_2)\,dx + \int_{R_2} P(x, c_1)\,dx$

② In the error made in R_1:

we always have $P(c_1|x) \geq P(c_2|x)$,

which implies $P(c_2|x) \leq \{P(c_1|x)P(c_2|x)\}^{\frac{1}{2}}$

$$\int_{R_1} P(x, c_2)\,dx = \int_{R_1} P(c_2|x)P(x)$$
$$\leq \int_{R_1} \{P(c_1|x)P(c_2|x)\}^{\frac{1}{2}} P(x)\,dx$$
$$= \int_{R_1} \{P(x,c_1)P(x,c_2)\}^{\frac{1}{2}} dx$$

③ In the error made in R_2:

we always have $P(c_2|x) \geq P(c_1|x)$,

which implies $P(c_1|x) \leq \{P(c_1|x)P(c_2|x)\}^{\frac{1}{2}}$

$$\int_{R_2} P(x, c_1)\,dx = \int_{R_2} P(c_1|x)P(x)\,dx$$
$$\leq \int_{R_2} \{P(c_1|x)P(c_2|x)\}^{\frac{1}{2}} P(x)\,dx$$
$$= \int_{R_2} \{P(x,c_1)P(x,c_2)\}^{\frac{1}{2}} dx$$

Substitute ②③ back to ①, we have

$$P(\text{mistake}) = \int_{R_1} P(x, c_2)\,dx + \int_{R_2} P(x,c_1)\,dx$$
$$\leq \int_{R_1} \{P(x,c_1)P(x,c_2)\}^{\frac{1}{2}} dx + \int_{R_2} \{P(x,c_1)P(x,c_2)\}^{\frac{1}{2}} dx$$
$$= \int \{P(x,c_1)P(x,c_2)\}^{\frac{1}{2}} dx \;\blacksquare$$

3.

(1) ① Assume $g(Y) = E[X|Y]$

$$\Rightarrow g(Y) = \sum_{x \in X} E[X = x \mid Y]\,P(x)$$
$$\Rightarrow g(Y) = \underset{x}{E}[x|Y]$$

② $E[X] = \sum_{y \in Y} E[X \mid Y = y]\,P(y)$

$$= \sum_{y \in Y} g(y)\,P(y)$$
$$= \underset{y}{E}[g(y)]$$
$$= \underset{y}{E}\left[\underset{x}{E}[x|y]\right] \;\blacksquare$$

(2) ① $\underset{x}{\mathrm{Var}}(x|Y) = \underset{x}{E}[x^2|Y] - \left(\underset{x}{E}[x|Y]\right)^2$

$$\Rightarrow \underset{y}{E}\left[\underset{x}{\mathrm{Var}}(x|y)\right] = \underline{\underset{y}{E}\left[\underset{x}{E}[x^2|y]\right]} - \underset{y}{E}\left[\left(\underset{x}{E}[x|y]\right)^2\right]$$
$$= \underline{E[x^2]} - \underset{y}{E}\left[\left(\underset{x}{E}[x|y]\right)^2\right]$$

② $\underset{y}{\mathrm{Var}}\left(\underset{x}{E}[x|y]\right) = \underset{y}{E}\left[\left(\underset{x}{E}[x|y]\right)^2\right] - \left(\underset{y}{E}\left[\underset{x}{E}[x|y]\right]\right)^2$

①② $\Rightarrow \underset{y}{E}\left[\underset{x}{\mathrm{Var}}(x|y)\right] + \underset{y}{\mathrm{Var}}\left(\underset{x}{E}[x|y]\right)$

$$= E[x^2] - \cancel{\underset{y}{E}\left[\left(\underset{x}{E}[x|y]\right)^2\right]} + \cancel{\underset{y}{E}\left[\left(\underset{x}{E}[x|y]\right)^2\right]} - \overbrace{\underline{\left(\underset{y}{E}\left[\underset{x}{E}[x|y]\right]\right)^2}}^{=(E[X])^2}$$
$$= E[x^2] - (E[x])^2 = \mathrm{Var}(x) \;\blacksquare$$