1. Linear Classification with a Margin

- Show that a dataset with two point is sufficient to determine the maximum-margin

Hint: $x_1 \in C_1 \; [y_1 = +1]$ $x_2 \in C_2 \; [y_2 = -1]$

- Setup minimization problem with constraints on $w^T x_i + b$... subject to:

$w^T x_1 + b = +1$ for optimization problem:

$w^T x_2 + b = -1$

$$\underset{w \in \mathbb{R}^p}{\operatorname{argmin}} \|w\|_2^2$$

- What is w? b?

Lagrange function and lagrange multipliers: $L(\lambda, w) = f(w) - \lambda g(x)$

$\lambda \geq 0$

$g(x_1) = w^T x_1 + b - 1$ $\quad g(x_2) = w^T x + b + 1$

$$L(\lambda_1, \lambda_2, w, b) = \operatorname{argmin}\left[\|w\|_2^2 + \lambda_1(w^T x_1 + b - 1) + \lambda_2(w^T x_2 + b + 1)\right]$$

$\|w\|_2^2 \rightarrow w^T w$

① Take derivatives:

$$\frac{\partial L}{\partial w} = 2w + \lambda_1 x_1 + \lambda_2 x_2 = 0 \; ; \; \frac{\partial L}{\partial b} = \lambda_1 + \lambda_2 = 0$$

② Find solution for terms:

$$\boxed{\lambda_1 = -\lambda_2}$$

$$0 = 2w + \lambda_1 x_1 + \lambda_2 x_2$$

$$= 2w - \lambda_2 x_1 + \lambda_2 x_2$$

$$= 2w + \lambda_2(x_1 - x_2)$$

$$-2w = -\lambda_2(x_1 - x_2)$$

$$\boxed{w = \frac{\lambda_2(x_1 - x_2)}{2}}$$

$w^T x_1 + b = 1 \qquad w^T x_2 + b = -1$

$w^T x_1 + b = -(w^T x_2 + b)$

$w^T x_1 + b = -w^T x_2 - b$

$2b = -w^T x_1 - w^T x_2$

$b = \frac{-w^T}{2}(x_1 + x_2)$

$$b = \frac{\left[\frac{\lambda_2}{2}(x_1 - x_2)\right]^T}{2}(x_1 + x_2)$$

$$\boxed{b = \frac{-\lambda_2}{4}(x_1 - x_2)^T(x_1 + x_2)}$$

Constraints:

Problem 2

- Find the result of minimizing the loss of sum of squared errors.
- Add a penalty for an L_2 penalty

$$\underset{w}{\operatorname{argmin}} \left\{ \sum (w^T x_i - y_i)^2 + \lambda \|w\|_2^2 \right\}$$

- How does this change the solution to the original regression solution?
- What is the impact of adding the penalty?

Logistic Function: $\frac{1}{1+\exp(-w^T x_i)}$

Cross Entropy: $L(w) = -\sum_{i=1}^{n} \{ y_i \log(\hat{y}_i) + (1+y_i) \log(\hat{y}_i) \}$

where $\hat{y} = \frac{1}{1+\exp(-w^T x)}$

$$\underset{w}{\operatorname{argmin}} \left\{ \sum \underbrace{(w^T x_i - y_i)^2}_{\downarrow} + \underbrace{\lambda \|w\|_2^2}_{\downarrow} \right\}$$

$(wx-y)(wx-y)^T \qquad \lambda w^T w$

① Rework equation

$$L(w) = (wx-y)^T(wx-y) + \lambda w^T w$$

$$= \bar{w}^T X^T \bar{w} X - \underbrace{\bar{w}^T X^T \bar{y} - \bar{y}^T \bar{w} x}_{2\bar{y}^T X \bar{w}} + \bar{y}^T \bar{y} + \lambda \bar{w}^T \bar{w}$$

$$L(w) = \bar{w}^T X^T \bar{w} X - 2\bar{y}^T X \bar{w} + \bar{y}^T \bar{y} + \lambda \bar{w}^T \bar{w}$$

$A^T \bar{x}^T \bar{b} + A \bar{x} b^T = 2b^T A x$

② Take derivative w.r.t w

$\frac{d}{dw}(2yX^T w) = 2\bar{y}^T X$

$\frac{d}{dw}(X^T \bar{w}^T X \bar{w}) = 2X^T X \bar{w}$

$$\frac{dL}{dw} = 2X^T X \bar{w} - 2\bar{y}^T X + 0 + 2\lambda w = 0$$

$$2(X^T X w - \bar{y}^T X + \lambda \bar{w}) = 0$$

$$\bar{w}(X^T X + \lambda I) - \bar{y}^T X = 0$$

$$w(X^TX + \lambda I) = \bar{y}^T X$$

$$\bar{w} = \frac{\bar{y}^T X}{(X^TX + \lambda I)}$$

③ New solution with penalty

$$L(w) = (\bar{w}^T X - y)(\bar{w}^T X - y)$$

$$\bar{w}^T X^T \bar{w} X + 2y^T X w + y^T y$$

④ Original solution without penalty

$$\frac{\partial L}{\partial w} = X^T X \bar{w} - y^T X = 0 \rightarrow \bar{w} = \frac{\bar{y}^T X}{X^T X}$$

The add penalty prevents $\bar{w}$ values from getting too high.

4.

- The Baye decision rule describes an approach we take for choosing a ω for x.
- Can be achieved by modeling $P(\omega|x)$, or $\frac{P(x|\omega)P(\omega)}{P(x)}$
- What are the Pro/Cons for each approach?
- Why is know $P(x)$ useful?

$P(\omega|x)$:- Directly estimate to probability data belongs to a class, however difficult to determine outliers.

$\frac{P(x|\omega)P(\omega)}{P(x)}$:- Obtain useful information of the data by finding the prior, likelihood, evidence, probabilities

$P(x)$: Since the likelihood $P(x|\omega)$ can be >1, $P(x)$ acts as a normalize factor to give an appropriate probability for the posterior