1. (a) training set $\{(x_i, y_i)\}_{i=1}^{n}$

$$l(w) = \sum_{i=1}^{n} l_i(w) = \sum_{i=1}^{n} -y_i \log h(x_i) - (1-y_i)\log(1-h(x_i))$$

$$h(x) = g(w^T x) = \frac{1}{1+\exp(-w^T x)}$$

$$P(y=1 \mid x; w) = h(x)$$

$$\nabla l(w) = \sum_{i=1}^{n} -y_i \cdot \nabla \log h(x_i) - (1-y_i)\nabla \log(1-h(x_i))$$

$$= \sum_{i=1}^{n} -y_i \frac{\nabla h(x_i)}{h(x_i)} - (1-y_i)\frac{1-\nabla h(x_i)}{1-h(x_i)}$$

$$\nabla h(x_i) = \nabla(1+\exp(-w^T x_i))^{-1} = -1\cdot(-x_i)\cdot\exp(-w^T x_i)\cdot(1+\exp(-w^T x_i))^{-2}.$$

$$= \frac{x_i \cdot \exp(-w^T x_i)}{(1+\exp(-w^T x_i))^2} = x_i \cdot h(x)^2 \cdot \exp(-w^T x_i)$$

$$\nabla l(w) = \sum_{i=1}^{n} -y_i \cdot \frac{x_i \cdot h(x_i)^2 \cdot \exp(-w^T x_i)}{h(x_i)} - (1-y_i)\cdot\frac{\cancel{\nabla}(1 - x_i h(x_i)^2 - \exp(-w^T x_i))}{1-h(x_i)}$$

$$= \sum_{i=1}^{n} -y_i\, x_i \,\cancel{h(x_i)}\, \frac{\exp(-w^T x_i)}{1+\exp(-w^T x_i)} - (1-y_i)\frac{x_i}{1+\exp(-w^T x_i)}.$$

$$= \sum_{i=1}^{n} -y_i x_i (1-h(x_i)) - (1-y_i) x_i \cdot h(x_i)$$

$$= \sum_{i=1}^{n} 2 y_i x_i h(x_i) - y_i x_i - x_i h.$$

(b) $\nabla^2(L(w)) = \nabla X^T h(x)$

$= XSX^T$, where S is diagonal.

$h(x_i) > 0$. and $\exp(-w^T x_i) > 0$.

it is positive semidefinite and $L(w)$ is convex

The global one.

2. (a) $f(x;\alpha,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\alpha)^2}{2\sigma^2}\right)$

$L(\theta) = \prod_{i=1} f(x_i;\theta)$

$\log L(\alpha;x) = n \cdot \log(2\pi\sigma^2)^{\frac{1}{2}} - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (x_i-\alpha)^2$

$\frac{\partial \log L(\alpha;x)}{\partial x} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i-\alpha) = 0$

$$\boxed{\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

$\frac{\partial \log L(\alpha;x)}{\partial(\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (x_i-\alpha)^2.$

$$\begin{cases} \frac{1}{\sigma^2} \sum (x_i-\hat{\alpha}) = 0. \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \cdot \sum (x_i-\alpha)^2 = 0 \end{cases}$$

$$\begin{cases} \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x-\hat{\alpha})^2. \\ \hat{\alpha} = \frac{1}{n} \sum_{i=1}^{\hat{}} x_i \end{cases}$$

b) $f(x;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$

$$L(\mu; x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \cdot \exp\left(-\frac{1}{2}\sum_{k=1}^{n}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)\right)$$

$x_k,\ k \in \{1, 2, \dots, n\}$

Let S denotes $S = \sum_{k=1}^{n}(x_k-\overline{x})(x_k-\overline{x})^T$.

$$L(\mu; x_1, \dots, x_n) = \frac{1}{(2\pi)^{np/2} \cdot |\Sigma|^{n/2}} \cdot \exp\left(-\frac{1}{2}tr(\Sigma^{-1}S) - \frac{1}{2}n(\overline{x}-\mu)^T\Sigma(\overline{x}-\mu)\right)$$

$$\log l(\mu; x_1, \dots, x_n) = \log\frac{1}{(2\pi)^{np/2} \cdot |\Sigma|^{\frac{n}{2}}} - \frac{n}{2}\log|\Sigma| - \frac{1}{2}tr(\Sigma^{-1}S) - \frac{1}{2}n(\overline{x}-\mu)^T\Sigma^{-1}(\overline{x}-\mu)$$

$$\frac{\partial \log l(\mu; x_1, \dots, x_n)}{\partial \mu} = 0.$$

$\therefore \overline{x} - \mu = 0$

$\therefore \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$

(a)

$$H(X) - H(X|Y) = -\int p(x)\ln(p(x))\,dx + \int p(x,y)\ln(x|y)\,dy\,dx$$

$$= -\int \left(\int p(x,y)\ln(x|y)\,dy\right) - p(x)\ln P(x)\,dx.$$

$$= -\int \left(\int p(x,y)\cdot(\ln P(y|x) + \ln P(x) - \ln P(y))\,dy\right) - P(x)\ln P(x)\,dx$$

$$= \int \left(\int p(x,y)\ln p(y|x)\,dy + \int p(x,y)\,dy\cdot\ln P(x) - \int p(x,y)\ln P(y)\,dy - p(x)\ln P(x)\right)dx$$

$$= -H(y|x) + \iint p(x,y)\ln\frac{P(x)}{P(y)} - P(x)\ln P(x)\,dy\,dx.$$

$$= -H(y|x) + \iint P(x,y)\ln P(y)\,dy\,dx$$

$$= -H(y|x) + \int P(y)\ln P(y)\,dy$$

$$= -H(y|x) + H(y)$$

$$= H(y) - H(y|x)$$

$$= I(x,y)$$

(b) $I(X,Y) = H(X) - H(X|Y)$

$$= -\int p(x) \ln p(x)\, dx + \int p(x,Y) \ln(x|Y)\, dx\, dY.$$

$$= -\int p(f(Y)) \ln p(f(Y))\, df(Y) + \int p(f(Y),Y) \underbrace{\ln p(f(Y)|Y)}_{=1}\, df(Y)\, d$$

$$= -\int p(Y) \ln p(Y)\, dY$$

$$= H(Y)$$

$$I(X,Y) = -\int p(f(Y)) \ln p(f(Y))\, d f(Y)$$

$$= H(f(Y))$$

$$= H(X)$$

$I(X,Y) = H(X) = H(Y)$

(C) $\hat{p}(x) \triangleq \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[x = x_i]$ — ①

$$\min_{\theta} D_{KL}(\hat{p} \| q) \triangleq \min_{\theta} - \int \hat{p}(x) \ln \frac{q(x|\theta)}{\hat{p}(x)} dx$$

$$= \min_{\theta} - \int \hat{p}(x) \ln q(x|\theta) dx + \int \hat{p}(x) \ln \hat{p}(x) dx.$$

$$\propto \min_{\theta} - \int \hat{p}(x) \ln q(x|\theta) dx.$$

plug in ① $\Rightarrow$ $= \min_{\theta} - \int \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[x = x_i] \ln q(x|\theta) dx$

$$= \min_{\theta} - \frac{1}{N}\sum_{i=1}^{N} \int \delta(x - x_i) \ln q(x|\theta) dx$$

$$= \min_{\theta} - \frac{1}{N}\sum_{i=1}^{N} \ln q(x|\theta) \propto \max_{\theta} q(x|\theta)$$

$\max_{\theta} q(x|\theta)$ is the maximum likelihood estimation given D

(d) objective: $\max \int_{-\infty}^{\infty} p(x) \ln p(x) \, dx$

constraints: $\begin{cases} \int_{-\infty}^{\infty} p(x) \, dx = 1 \\ \int_{-\infty}^{\infty} x p(x) \, dx = \mu \\ \int_{-\infty}^{\infty} (x-\mu)^2 p(x) \, dx = \delta^2 \end{cases}$

$$-\int_{-\infty}^{\infty} p(x) \ln p(x) \, dx + \lambda_1 \left(\int_{-\infty}^{\infty} p(x) \, dx - 1 \right) + \lambda_2 \left(\int_{-\infty}^{\infty} x p(x) \, dx - \mu \right)$$
$$+ \lambda_3 \left(\int_{-\infty}^{\infty} (x-\mu)^2 p(x) \, dx - \delta^2 \right) = F(p(x))$$

$$\frac{\partial F(p(x))}{\partial x} = 0$$

$$\Rightarrow \hat{p}(x) = \frac{1}{\sqrt{2\pi}\delta} \exp\left(-\frac{(x-\mu)^2}{2\delta^2} \right)$$

4. (a) $w = \operatorname{argmin} \sum c_i (y_i - \beta^T x_i - b)^2$

$= \operatorname{argmin} \| C^{\frac{1}{2}} (y - Xw) \|_2^2$

where C is diagonal.

$$\frac{\partial L(w)}{\partial w} = -2 (C^{\frac{1}{2}} X)^T C^{\frac{1}{2}} (y - Xw) = 0$$

$$\hat{w} = ((CX)^T X)^{-1} (CX)^T y = (X^T C X)^{-1} X^T C y.$$

$c_i = 1$, $\hat{w} = (X^T X)^{-1} X^T y$, $\hat{w} = \begin{bmatrix} b \\ \beta \end{bmatrix}$

$y_i = \beta^T x_i + b + \varepsilon_i$

$y = Xw + \varepsilon$ where $\varepsilon | X \sim N(0, \sigma^2 I)$

$$\max P(y|w, X) = \max - (y - Xw)^T (\sigma^2 I)^{-1} (y - Xw)$$

$$= \min_w \| y - Xw \|_2^2$$

$$\hat{w}_{ML} = (X^T X)^{-1} X^T y = \hat{w}_{LS}$$

(b) $y|X,w \sim N(Xw, \Sigma)$

$$\max P(y|X,w) = \min (y-Xw)^T \Sigma^{-1} (y-Xw)$$

$$= \min \|\Sigma^{-\frac{1}{2}}(y-Xw)\|_2^2$$

$$\hat{w}_{LS} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} y.$$

The MLE of w with different noise variance for each i is equivalant to weight LS. with matric $C = \Sigma^{-1}$

5) 1a) $\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{N} \varepsilon_i$

Subject to $t^{(i)}(w^T x^{(i)}+b) \geq 1-\varepsilon_i$, where $\varepsilon_i \geq 0$

$$S_i \geq \max\left[0, 1-t^{(i)}(\underline{w}^T x^{(i)}+b)\right]$$

So $\min\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} S_i\right)$ is equivalent to the

~~$S_i \geq \max$~~ $\min \frac{1}{2}\|w\|_2^2 + C\sum \max\left(0, 1-t^{(i)}(\underline{w}^T x^{(i)}+b)\right)$

1b) $\rho(w,b) = \min d(x_i, H) = \min \frac{|w^T x_i + b|}{\|w\|_2}$.

scale w and b by $\frac{1}{\min|w^T x_i + b|}$,

$\min \rho(\hat{w}, \hat{b}) = \frac{1}{\|\hat{w}\|_2}$

$y_i(|w^T x_i + b|) = 1. \Rightarrow \rho_{min} = \frac{n}{\|w\|_2}$. where $n = \min|w^T x_i + b|$.

$\rho_i = \frac{|w^T x_i + b|}{\|w\|_2}$ $\forall i$ $t^{(i)}(w^T x^{(i)} + b) \geq 1-\varepsilon_i$

$\therefore \frac{\rho_i}{\rho_{min}} = \varepsilon_i^*$ and $\rho_i \propto \varepsilon_i^*$

(c)

$$\min\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\max\left(0, 1-t^{(i)}(w^T x^{(i)}+b)\right)\right) \quad \cdots \; ①$$

when $C \to \infty$, ① is equivalant to $\min\left(\max\left(0, 1-t^{(i)}(w^T x^{(i)}+b)\right)\right)$

$t^{(i)}(w^T x^{(i)}+b) = 1 \iff x^{(i)}$ is the closest point to the margin.

The SVM hard margin

$\min \frac{1}{2}\|w\|_2^2$ s.t. the closest point to the margin

these two are equivalant.