

Garnett
Menz

① Let $l(w) = \sum_{i=1}^n l_i(w) = \sum_{i=1}^n -y_i \log h(x_i) - (1-y_i) \log(1-h(x_i))$
 where $h(x) = g(w^T x) = \frac{1}{1+e^{-w^T x}}$.

a) Find $\nabla l(w)$.

$$\nabla l(w) = \left(\frac{\partial l(w)}{\partial w_1}, \frac{\partial l(w)}{\partial w_2}, \dots, \frac{\partial l(w)}{\partial w_n} \right).$$

$$\begin{aligned} \text{So } \frac{\partial l(w)}{\partial w_i} &= \frac{\partial l_i(w)}{\partial w_i} = \sum_{i=1}^n \frac{-y_i}{h(x_i)} \frac{\partial h(x_i)}{\partial w_i} - \frac{(1-y_i)}{1-h(x_i)} \left(-\frac{\partial h(x_i)}{\partial w_i} \right) \\ &= \sum_{i=1}^n \left(\frac{-y_i}{h(x_i)} + \frac{(1-y_i)}{1-h(x_i)} \right) \left(\frac{\partial h(x_i)}{\partial w_i} \right) \\ &= \sum_{i=1}^n \frac{(h(x_i)-1)y_i + (1-y_i)h(x_i)}{h(x_i)(1-h(x_i))} \left(\frac{\partial h(x_i)}{\partial w_i} \right) \\ &= \sum_{i=1}^n \frac{h(x_i) - y_i}{h(x_i)(1-h(x_i))} \left(\frac{\partial h(x_i)}{\partial w_i} \right) \end{aligned}$$

$$\text{let } x_i = (x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}) = x^{(i)}$$

Then we find $\frac{\partial h(x_i)}{\partial w_j} = \frac{\partial}{\partial w_j} \left(\frac{1}{1+e^{-w^T x^{(i)}}} \right)^{-1}$

$$= \frac{-1}{(1+e^{-w^T x^{(i)}})^2} - x_j^{(i)} e^{-w^T x^{(i)}} = \frac{x_j^{(i)} e^{-w^T x^{(i)}}}{(1+e^{-w^T x^{(i)}})^2} = x_j^{(i)} (h(x^{(i)})(1-h(x^{(i)})))$$

so

$$\nabla l(w) = \sum_{i=1}^n x_j^{(i)} (h(x^{(i)}) - y^{(i)}).$$

□

b) Find H for $l(w)$, and show that it is PSD, and thus that $l(w)$ is convex.

$$H_{kj} = \frac{\partial^2 l}{\partial w_k \partial w_j} = \frac{\partial}{\partial w_k} \left(\sum_{i=1}^n x_j^{(i)} (h(x^{(i)}) - y^{(i)}) \right) \\ = \sum_{i=1}^n x_j^{(i)} \frac{\partial h(x^{(i)})}{\partial w_k} = \sum_{i=1}^n x_j^{(i)} x_k^{(i)} (h(x^{(i)}) (1 - h(x^{(i)})))$$

$$\text{So } H = \sum_{i=1}^n x^{(i)} (x^{(i)})^T h(x^{(i)}) (1 - h(x^{(i)}))$$

(if we assume $x^{(i)}$ is a column vector).

We show H is PSD. Let A be a column vector in \mathbb{R}_+^n . Then

$$A^T H A = \sum_{i=1}^n A^T x^{(i)} (x^{(i)})^T A \cdot h(x^{(i)}) (1 - h(x^{(i)})) \\ = \sum_{i=1}^n (A \cdot x^{(i)})^2 h(x^{(i)}) (1 - h(x^{(i)})) \\ = \sum_{i=1}^n \frac{(A \cdot x^{(i)})^2}{(1 + e^{-w^T x^{(i)}})^2}$$

This term is $\geq 0 \forall A \in \mathbb{R}^n$ ($A \neq 0$)

So H is positive semidefinite, and thus, $l(w)$ is convex.

□

c) Use this on the Classification data set:

See attached code.

② Let $L(\theta) = \prod_{i=1}^n f(x_i; \theta)$ be a likelihood, and

$$l(\theta) = \sum_{i=1}^n \log f(x_i; \theta).$$

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

Find the MLE for α, σ^2 :

$$\begin{aligned} \log(f(x_i; \theta)) &= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x_i-\alpha)^2}{2\sigma^2} \\ &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x_i-\alpha)^2}{2\sigma^2} \\ &= -\frac{1}{2} \log(2\pi) - \log(\sigma) - \frac{(x_i-\alpha)^2}{2\sigma^2} \end{aligned}$$

$$\text{so } l_{\alpha, \sigma} = -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \sum_{i=1}^n \frac{(x_i-\alpha)^2}{2\sigma^2}$$

$$\text{Maximize: } \frac{dl}{d\sigma} = -\frac{n}{\sigma} + \sum_{i=1}^n \frac{2(x_i-\alpha)^2}{2\sigma^3} = \frac{1}{\sigma} \left(-n + \frac{1}{\sigma^2} \sum_{i=1}^n (x_i-\alpha)^2 \right) = 0$$

$$\text{so } \frac{1}{\sigma^2} \sum_{i=1}^n (x_i-\alpha)^2 = n, \quad \text{i.e.,}$$

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i-\alpha)^2}{n}$$

$$\text{Similarly, } \frac{\partial l}{\partial \alpha} = -\sum_{i=1}^n \frac{2(x_i - \alpha)}{2\sigma^2} = -\sum_{i=1}^n \frac{(x_i - \alpha)}{\sigma^2}.$$

So $\frac{\partial l}{\partial \alpha} = 0$ gives $(\sum_{i=1}^n x_i) - n\alpha = 0$ i.e.,

$$\alpha = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}.$$

So putting it all together,

$$\alpha = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}, \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

b) Let $f(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))$. Find MLE for μ, Σ . □

$$\text{Then, } l(\theta) = \sum_{i=1}^n -\frac{1}{2} \log((2\pi)^d |\Sigma|) - \frac{1}{2} (x^{(i)} - \mu)^T \Sigma^{-1} (x^{(i)} - \mu)$$

$$= \sum_{i=1}^n -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma|) - \frac{1}{2} (x^{(i)} - \mu)^T \Sigma^{-1} (x^{(i)} - \mu)$$

$$= \sum_{i=1}^n \left(\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma|) - \frac{1}{2} \sum_{j=1}^m \sum_{a=1}^m (x_a^{(i)} - \mu_a) \Sigma_{aj}^{-1} (x_j^{(i)} - \mu_j) \right)$$

$$\text{So } \frac{\partial l}{\partial \mu_k} = \sum_{i=1}^n \left(\frac{1}{2} \sum_{j \neq k} \Sigma_{kj}^{-1} (x_j^{(i)} - \mu_j) + \sum_{a \neq k} (x_a^{(i)} - \mu_a) \Sigma_{ak}^{-1} \right. \\ \left. + 2 \sum_{kk}^{-1} (x_k^{(i)} - \mu_k) \right). \text{ But } \Sigma \text{ is symmetric, so } \Sigma^{-1} \text{ is too, so}$$

$$\frac{\partial l}{\partial \mu_k} = \sum_{i=1}^n \sum_{a=1}^m \sum_{ak}^{-1} (x_a^{(i)} - \mu_a) = \sum_{i=1}^n (\sum_{a=1}^m (x_a^{(i)} - \mu_a))_k =$$

Setting this equal to zero $\forall k$, we get

$$\sum_{i=1}^n (\sum_{a=1}^m x_a^{(i)}) = n \sum_{a=1}^m \mu_a. \text{ Multiply by } \Sigma:$$

$$\sum_{i=1}^n x^{(i)} = n\mu, \text{ so } \mu = \frac{1}{n} \sum_{i=1}^n x^{(i)}. \quad \square$$

③ a) Show that $I(x, y) = H(x) - H(x|y) = H(y) - H(y|x)$

where $I(x, y) = \int p(x, y) \ln\left(\frac{p(x, y)}{p(x)p(y)}\right) dx dy$

$$H(x) = -\int p(x) \ln(p(x)) dx$$

and $H(x|y) = -\int p(x|y) \ln(p(x|y)) dx dy$.

we have

$$I(x, y) = \int p(x, y) \ln\left(\frac{p(x, y)}{p(x)p(y)}\right) dx dy - \int p(x, y) \ln(p(x)) dx dy$$

But $p(x|y) = \frac{p(x, y)}{p(y)}$, so this is

$$\begin{aligned} I(x, y) &= \int p(x|y) p(y) \ln(p(x|y)) dx dy - \int p(x, y) \ln(p(x)) dx dy \\ &= - \int dy p(y) \left(\int p(x|y) \ln(p(x|y)) dx \right) - \int dx \ln(p(x)) \left(\int dy p(x, y) \right) \\ &\quad - \int dy p(y) H(x|y=y) - \int p(x) \ln(p(x)) dx \\ &= -H(x|y) + H(x) = H(x) - H(x|y). \end{aligned}$$

Because I is symmetric in X and Y , we see that we can follow the same logic to get

$$I(x, y) = H(y) - H(y|x).$$

□

b) Prove that if X and Y are related by a bijection (i.e., if $X = f(Y)$ and $Y = f^{-1}(X)$),
 $I(x,y) = H(x) = H(y)$.

If $X = f(Y)$, then $P(X|Y) = 1$ and $P(Y|X) = 1$ (i.e., knowing X means we know Y and vice versa).

$$\text{Thus, } H(x|y) = - \int 1 \ln(1) dy dx = - \int 0 = 0$$

$$\text{So } I(x,y) = H(x) = H(y).$$

c) Suppose we observe N samples $D = (x_1, \dots, x_N)$ from some unknown distribution. Define $\hat{p}(x)$ to be the empirical probability density estimate:

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^N I(x=x_i)$$

Let $g(x|\theta)$ be the prob. density corresponding to some known model with parameter θ .

Show that the nn K-L divergence

$$\min_{\theta} D_{KL}(p||g) \stackrel{\Delta}{=} \min_{\theta} \left(- \int \hat{p}(x) \ln \left(\frac{g(x|\theta)}{\hat{p}(x)} \right) dx \right)$$

is obtained by θ_m given the data D .

$$\text{we see that } D_{KL}(p||g) = \int \hat{p}(x) \ln \left(\frac{\hat{p}(x)}{g(x|\theta)} \right) dx$$

$$= \int \hat{p}(x) \ln(\hat{p}(x)) dx - \int \hat{p}(x) \ln(g(x|\theta)) dx$$

$$= H(x) - \int \hat{p}(x) \ln(g(x|\theta)) dx$$

$$= H(x) - \frac{1}{N} \int \sum_{i=1}^N \ln(g(x_i|\theta)) \cdot I(x=x_i) dx$$

$$= H(x) - 1/N \ln(g(x_i|\theta))$$

Thus, we see that

$$\begin{aligned}\underset{\theta}{\operatorname{argmax}}(D_{KL}(p||g)) &= \underset{\theta}{\operatorname{argmin}}(H(g) - \frac{1}{n} \sum g(x; \theta)) \\ &= \underset{\theta}{\operatorname{argmax}}(\frac{1}{n} \sum g(x; \theta))\end{aligned}$$

But this is just maximizing the likelihood for g , so we are done!

- d) Let $p = N(\mu, \sigma^2)$ be a Gaussian distribution and let g be any probability density with mean μ and variance σ^2 . Prove that $H(g) \leq H(p)$, that is, the Gaussian has maximum entropy among all equivariant distributions.

We maximize a new function:

$$\begin{aligned}J(p) &= - \int p(x) h_1(p(x)) dx + \lambda_0 (\int p(x) dx - 1) \\ &\quad + \lambda_1 (\int x p(x) dx - \mu) + \lambda_2 (\int (x - \mu)^2 p(x) dx - \sigma^2)\end{aligned}$$

(i.e., maximize H subject to the constraints that p is a pdf with mean μ and variance σ^2).

Then $\frac{dJ}{dp} = -\frac{p(x)}{p(x)} - h(p(x)) + \lambda_0 + \lambda_1 x + \lambda_2 (x - \mu)^2 = 0$

i.e., $p(x) = e^{-1 + \lambda_0 + \lambda_1 x + \lambda_2 (x - \mu)^2}$

Then the constraint equations:

$$\int p(x) dx = 1, \quad \int x p(x) dx = \mu, \text{ and}$$

$$\int (x - \mu)^2 p(x) dx = \sigma^2$$

give:

(Solving with Mathematica)

$$\lambda_1 = 0, \lambda_2 = 1 + \ln\left(\frac{1}{\sigma^2}\right) \text{ and } \lambda_3 = -\frac{1}{\sigma^2}.$$

Thus,

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ as expected. } \square$$

(4) Weighted Least Squares

a) Let $w = [b \ \beta^T]^T = \underset{\beta, b}{\operatorname{argmin}} L(\beta, b) = \underset{\beta, b}{\operatorname{argmin}} \sum_{i=1}^n c_i (y_i - \beta^T x_i - b)^2$

Find a closed-form solution for w in terms of X and C .

Let $X = \begin{pmatrix} 1 & x_1 \\ & x_2 \\ & \vdots \\ & x_n \end{pmatrix}$, $w^T = (b, \beta^T)$, and let y, C
let C be the square matrices with y_i, c_i on the diagonals.

Then

$$\begin{aligned} L(\beta, b) &= \sum_{i=1}^n c_i (y_i - w^T x_i)^2 = (y - w^T X)^T C (y - w^T X) \\ &= (y^T - (X^T w)) C (y - w^T X) = (y^T C y - x^T w C y - y^T C w^T X + x^T w C (w^T X)^T) \end{aligned}$$

$$\text{So } \frac{dL}{dw} = 2(-X^T C y + X^T C X w)$$

$$\text{So } \frac{dL}{dw} = 0 \text{ gives}$$

$$(X^T C X) w = (X^T C y) \quad \text{i.e.,}$$

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

Next, model $y_i = \beta^T x_i + b + \epsilon_i \quad \forall i$

The $y_i - \beta^T x_i - b$ is normally distributed
about 0 with variance σ^2 .

Suppose now we wish to do a likelihood fit.
Because we assume equal variance,

$$L(\beta, b) = e^{-\frac{(y_1 - \beta^T x_1 - b)^2}{2\sigma^2}} e^{-\frac{(y_2 - \beta^T x_2 - b)^2}{2\sigma^2}} \dots$$

$$\text{so } \log L = -\sum_{i=1}^n \frac{(y_i - \beta^T x_i - b)^2}{2\sigma^2} = \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta^T x_i - b)^2,$$

and minimizing $\log L$ is the same as maximizing $\sum_{i=1}^n (y_i - \beta^T x_i - b)^2$, that is, performing a least-squares fit.

b) Repeat part a), but with $y_i = \beta^T x_i + b + \epsilon_i$, where $\epsilon_i | x_i \sim N(0, \sigma_i^2)$.

We see that our value of w remains the same, but we must redefine C_i : As in (a),

$$L(\beta, b) = e^{-\frac{(y_1 - \beta^T x_1 - b)^2}{2\sigma_1^2}} e^{-\frac{(y_2 - \beta^T x_2 - b)^2}{2\sigma_2^2}} \dots \text{etc}$$

$$\text{Then } \log L = -\sum_{i=1}^n \frac{(y_i - \beta^T x_i - b)^2}{2\sigma_i^2} = -\sum_{i=1}^n C_i (y_i - \beta^T x_i - b)^2$$

$$\text{where } C_i = \frac{1}{2\sigma_i^2}.$$

Then, as in (a), maximizing $\log L$ is the same as minimizing $\sum_{i=1}^n C_i (y_i - \beta^T x_i - b)^2$. □

⑤ a) Show that the minimization problems

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i$$

$$\begin{cases} \xi_i \geq 0 \\ \text{subject to } t^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i \\ (i=1, \dots, N) \end{cases}$$

and

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \max(0, 1 - t^{(i)}(w^T x^{(i)} + b))$$

We note that $0 \leq \xi_i$ and $(1 - t^{(i)}(w^T x^{(i)} + b)) \leq \xi_i$.

Since we are minimizing the ξ_i , we see that at the solution, we have equality: either $\xi_i = (1 - t^{(i)}(w^T x^{(i)} + b)) \geq 0$ or $\xi_i = 0$ and $(1 - t^{(i)}(w^T x^{(i)} + b)) \leq 0$.

$$\text{Thus, } \xi_i = \max(0, 1 - t^{(i)}(w^T x^{(i)} + b))$$

So the problem becomes

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \max(0, 1 - t^{(i)}(w^T x^{(i)} + b)).$$

b) Let (w^*, b^*, ξ^*) be the solution of the second problem. Show that if $\xi_i^* > 0$, then the distance from $x^{(i)}$ to $t^{(i)}(w^*{}^T x^{(i)} + b^*) = 1$ is proportional to $-\xi_i^*$.

$$\text{If } \xi_i^* > 0, \quad \xi_i^* = (1 - t^{(i)}(w^*{}^T x^{(i)} + b^*))$$

Lemma: The distance from $x^{(i)}$ to the hyperplane $w^T x + b = 0$ is found by minimizing the Lagrangian

$$L = \|x - x^{(i)}\|^2 + 2\lambda(w^T x + b)$$

(let us use a multiplier of 2λ)

$$= \|x\|^2 - 2(x^{(i)} - \lambda w)^T x + 2\lambda b + \|x^{(i)}\|^2$$

$$\text{So } \frac{\partial L}{\partial x} = -2(x^{(i)} - \lambda w) + 2x$$

and $\frac{\partial L}{\partial x} = 0$ gives
 $x = (x^{(i)} - \lambda w)$

So we have $L = \sum w^T(x^{(i)} - \lambda w) + b = 0 \}$
 i.e., $\sum w^T x^{(i)} + b - \lambda \|w\|^2 = 0 \}$

$$\text{So } \lambda = \frac{w^T x^{(i)} + b}{\|w\|^2}$$

So the shortest distance is
 $\|x - x^{(i)}\| = \|\lambda w\| = \frac{\|w^T x^{(i)} + b\|}{\|w\|^2}$. \square

With this in mind, the distance from $x^{(i)}$ to
 $1 - t^{(i)}(w^* x + b^*) = 0$ is

$$d = \frac{\|t^{(i)} w^T x^{(i)} + 1 - t^{(i)} b^*\|}{\|t^{(i)} w\|^2} =$$

$$= \frac{\|s_i^*\|}{\|t^{(i)} w\|^2}, \text{ i.e., } d \text{ is proportional to } s_i.$$

c) Show that taking $C \rightarrow \infty$ gives the hard-margin SVM.

We see that taking $C = \infty$ enforces $s_i = 0 \forall i$, (otherwise, we cannot minimize anything). Thus,

$$1 - t^{(i)}(w^T x^{(i)} + b) \leq 0, \text{ i.e.,}$$

$t^{(i)}(w^T x^{(i)} + b) \geq 1$. This, however, is the hard-margin constraint, so we have shown that the hard-margin problem is equivalent to taking $C \rightarrow \infty$. \square