1、

1. False Logistic regression does not make assumption of how data was generated.

2. False $\log \frac{p}{1-p} = X\beta$, we can replace X with $\phi(X)$

3. False $\hat{y} = \frac{e^{X\beta}}{1+e^{X\beta}}$ can only fall between 0 to 1. It cannot estimate any y out of this range. So regression usually does not work in $\mathbb{R}$

4. False. Hard-margin does not have a solution when data is not linearly separable.

5. True; support vectors determine the maximum margin and hold hyperplane $f(x) = w^T x + b$

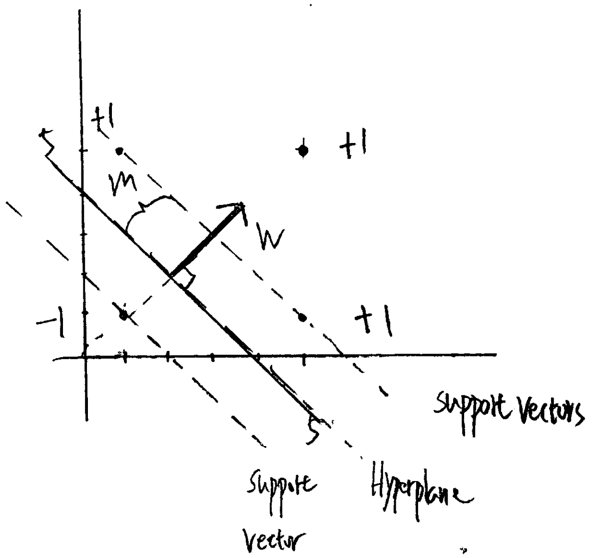
6. False, as C increases, ε_i are forced to decrease, and therefore lesser points are allowed to be inside the margin.

7. False, as C approaches to 0, more points are allowed to fall in margin. It is when $C \to \infty$, the problem becomes hard-margin.

8. False. Suppose C is a constant, the new formulation is more strict than regular SVM. It requires $\varepsilon_i = \varepsilon_j$ for some i and j. While the regular soft-margin SVM does not require this. Every data point can have its own slack variable.
   Suppose $(i,j) \in \{1, \cdots, n\}^2$, x_i and x_j are not on the margin, then $\varepsilon_i = \varepsilon_j = 0$
   The objective value $\|w\|_2^2 + C\sum_{i=1}^{n}\varepsilon_i$ of new formulation is the same as the old one.

2. SVM

(1)



(2) $W = C \cdot (1, 1)^T$

C is a constant

$f(x) = x^T w + b = 0$ defines hyperplane

(3) $\frac{|W^T(Z - X_0)|}{\|W\|_2} = D = \frac{|W^T Z - W^T X_0|}{\|W\|_2} = \frac{|W^T Z + b|}{\|W\|_2}$ it's easy to calulate the margin m from the graph.

$\frac{2}{\|W\|_2} = 2m \quad \|W\|_2 = \frac{\sqrt{2}}{2}$

$m = \sqrt{2}$

(4) $\sqrt{C^2 + C^2} = \frac{\sqrt{2}}{2} \quad C = \frac{1}{2}$

$W = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$

$\frac{-1 \cdot (\frac{1}{2} + \frac{1}{2} + b)}{\frac{\sqrt{2}}{2}} = \sqrt{2} \quad b = -2$

verify on (5, 1) $\frac{1 \cdot (\frac{5}{2} + \frac{1}{2} + b)}{\frac{\sqrt{2}}{2}} = \sqrt{2} \quad b = -2$

3.1. follow the definition,

$$f(z) \geq \underbrace{f_K(z) \geq f_K(w) + g_w^T(z-w)}_{\text{the definition}}$$

replace $f_K(w) = f(w)$

$$f(z) \geq f(w) + g_w^T(z-w)$$

Therefore, g_w is also a subgridiene of f at w.

3.2. if $1 - yw^Tx > 0 \Rightarrow yw^Tx < 1$

the subgridiene of $f(w)$ is $-yx$

if $1 - yw^Tx < 0$

$$g_w = 0$$

if $1 - yw^Tx = 0$

$$g_w \in [-yx, 0]$$

in one line $\partial f(w) = \underset{\substack{\downarrow \\ \text{Indicator}}}{I}(y \cdot (w \cdot x) \leq 1)(-yx)$

3.3. the margin $w^T x_i + b > 0 \quad \hat{y}_i = 1$

$w^T x_i + b < 0 \quad \hat{y}_i = -1$

if $\hat{y}\, y$ agrees, $\ell(\hat{y}, y) = 0$ no penalty. if not agree, $\ell(\hat{y}, y) = -\hat{y}y > 0$ has penalty

if $\{x \mid \hat{w}^T x = 0\}$ is a separating hyperplane. It captures mismatch and thus has meaning.

Then D is separated by the hyperplane, which means there must be no mismatch.

So for every $\{x_i, y_i\}$ $\ell(x_i^T w, y_i) = 0$, then $L(w; D) = 0$

3.4 derive subgradiene first

$$\nabla_w \ell(x_i^T w, y_i) = \nabla_w \max\{0, -y_i w^T x_i\} = \begin{cases} -y_i x_i & \text{if } -y_i w^T x_i > 0 \\ 0 & \text{if } -y_i w^T x_i < 0 \\ [-y_i x_i, 0] \text{ eg } -y_i x_i & \text{if } -y_i w^T x_i = 0 \end{cases}$$

in one line

$$= I(y_i w^T x_i \leq 0)(-y_i x_i)$$

In Algorithm 1

for $i = 1, 2, \cdots, n$:

if $y_i x_i w^{(k)} \leq 0$ then

$w^{(k+1)} = w^{(k)} + y_i x_i$

else

$w^{(k+1)} = w^{(k)}$.

which is equivalene to SSGD wich $a = 1$

$$w^{(k+1)} = w^{(k)} - 1 \cdot g_k = w^{(k)} + I(y_i x_i w^{(k)} \leq 0)\, y_i x_i$$

5. $\hat{w} = w^{(n)} = w^{(n-1)} + I(y_i x_i w^{(n-1)} \leq 0) y_i x_i$

for every step, $w^{(k)}$ might be updated or not be updated

$\vdots$

$w^{(1)} = w^{(0)} + I(y_1 x_1 w^{(0)} \leq 0) y_1 x_1$ let $w^{(0)}$ be any $y_i x_i$

for convinience, $w^{(0)} = y_1 x_1$.

$\hat{w} = \sum_{i=1}^{n} a_i x_i$,

$a_1 = y_1$, $a_i = 0$ if $y_i x_i w^{(i-1)} > 0$ for $i = 2, \cdots, n$

$a_1 = 2y_1$, $a_i = y_i$ if $y_i x_i w^{(i-1)} \leq 0$

6. from 3.2 we have

for $y_i w^T x_i < 1$, $g_w = \lambda w - y_i x_i$

for $y_i w^T x_i > 1$, $g_w = \lambda w$

for $y_i w^T x_i = 1$, ∂_w of hinge loss can be any value in $[-y_i x_i, 0]$

take 0 in this case

then $g_w = \lambda w$

In conclusion, $g_w = \begin{cases} \lambda w - y_i x_i & \text{for } y_i w^T x_i < 1 \\ \lambda w & \text{for } y_i w^T x_i \geq 1 \end{cases}$

7. with SSGD, if $y_i w^T x_i < 1$

$$W^{(k+1)} = w^{(k)} - \alpha g_k$$

$$= w^{(k)} - \alpha \cdot (\lambda w^{(k)} - y_i x_i)$$

$$= w^{(k)} - \alpha \lambda w^{(k)} + \alpha y_i x_i$$

$$= w^{(k)} - \frac{1}{k} w^{(k)} + \frac{1}{k\lambda} y_i x_i$$

for Pegasos plug in $\eta = \frac{1}{k\lambda}$

$$w^{(k+1)} = (1 - \frac{1}{k}) w^{(k)} + \frac{1}{k\lambda} y_i x_i$$

$$= w^{(k)} - \frac{1}{k} w^{(k)} + \frac{1}{k\lambda} y_i x_i$$

if $y_i w^T x_i \geq 1$

$$W^{(k+1)} = w^{(k)} - \alpha \lambda w^{(k)}$$

$$= w^{(k)} - \frac{1}{k} w^{(k)}$$

for pegasos

$$w^{(k+1)} = (1 - \frac{1}{k}) w^{(k)}$$

$$= w^{(k)} - \frac{1}{k} w^{(k)}$$

We showed that the two methods are exactly the same.

4.1

likelihood function is : $lik(\beta) = \prod_{i=1}^{n} P_i^{y_i}(1-P_i)^{1-y_i}$ $\qquad P_i = \mathbb{P}(y_i=1 \mid X_i;\beta) = \frac{e^{X\beta}}{1+e^{X\beta}}$

take log $\ell(\beta) = \sum_{i=1}^{n} y_i \log(P_i) + (1-y_i)\log(1-P_i)$ $\qquad 1-P_i = \frac{1}{1+e^{X\beta}}$

$$= \sum_{i=1}^{n} y_i(X\beta - \log(1+e^{X\beta})) + (-\log(1+e^{X\beta})) - y_i(-\log(1+e^{X\beta}))$$

$$= \sum_{i=1}^{n} y_i X\beta - \log(1+e^{X\beta})$$

negative: $L(\beta) = \sum_{i=1}^{n} -y_i X\beta + \log(1+e^{X\beta})$

$$\nabla_\beta L(\beta) = -YX^T + \frac{e^{X\beta}\cdot X^T}{1+e^{X\beta}} = X^T\left(\frac{e^{X\beta}}{1+e^{X\beta}} - Y\right)$$

$$= X^T(P-Y)$$

$P = (P_1, \cdots, P_n)^T$

$Y = (y_1, \cdots, y_n)^T$

$$= \sum_{i=1}^{n} (P_i - y_i)(1, X_{i1}, \cdots, X_{ip})^T$$

Hessian

$$\nabla_\beta^2 L(\beta) = \nabla_\beta \sum_{i=1}^{n} P_i(1, X_{i1}, \cdots, X_{ip})^T$$

$$= \sum_{i=1}^{n} P_i(1-P_i)(1, X_{i1}, \cdots, X_{ip})^T(1, X_{i1}, \cdots, X_{ip})$$

$$= X^T W X$$

$$W = \begin{array}{cc} n & \begin{pmatrix} P_1(1-P_1) & & 0 \\ & \ddots & \\ 0 & & P_n(1-P_n) \end{pmatrix} \\ & n \end{array}$$ $n \times n$ diagonal matrix

4.2 Hessian

$$HL(\beta) = X^T W X$$

let z be any vector $\in \mathbb{R}^n \neq 0$

$$z^T X^T W X z = (Xz)^T W (Xz)$$ let $Xz = s$

$$= s^T \sqrt{W}^T \sqrt{W} s$$ W is diagonal so, $W = \sqrt{W} \cdot \sqrt{W}$

$$= \|\sqrt{W} s\|_2^2 \geq 0$$

Therefore $HL(\beta)$ is PSD, and $L(\beta)$ is convex.

We have a theorem: A twice differentiable function is convex iff its hessian is PSD.

4.3. Taylor's expansion of $L(\beta)$ around $\beta^{(m)}$

$$\tilde{L}(\beta) = L(\beta^{(m)}) + \nabla L(\beta^{(m)})^T (\beta - \beta^{(m)}) + \frac{(\beta-\beta^{(m)})^T HL(\beta^{(m)}) (\beta - \beta^{(m)})}{2}$$

$\tilde{L}(\beta) \simeq L(\beta)$ instead of minimizing $L(\beta)$ we minimizing $\tilde{L}(\beta)$

$$\nabla \tilde{L}(\beta) = \nabla L(\beta^{(m)}) + HL(\beta)(\beta - \beta^{(m)})$$

$\nabla \tilde{L}(\beta) = 0$ if and only if

$$HL(\beta)(\beta - \beta^{(m)}) = -\nabla L(\beta^{(m)})$$

$$\beta^{(m+1)} = \beta^{(m)} - HL(\beta)^{-1} \nabla L(\beta^{(m)})$$

4.4 plug in Hessian and $\nabla L(\beta)$

$$\begin{aligned}\beta^{(m+1)} &= \beta^{(m)} - (X^T W X)^{-1} \cdot X^T (P - Y) \\ &= (X^T W X)^{-1} X^T W X \cdot \beta^{(m)} - (X^T W X)^{-1} X^T W W^{-1} (P - Y) \\ &= (X^T W X)^{-1} X^T W Z \quad \text{where } Z = (X\beta^{(m)} - W^{-1}(P-Y))\end{aligned}$$

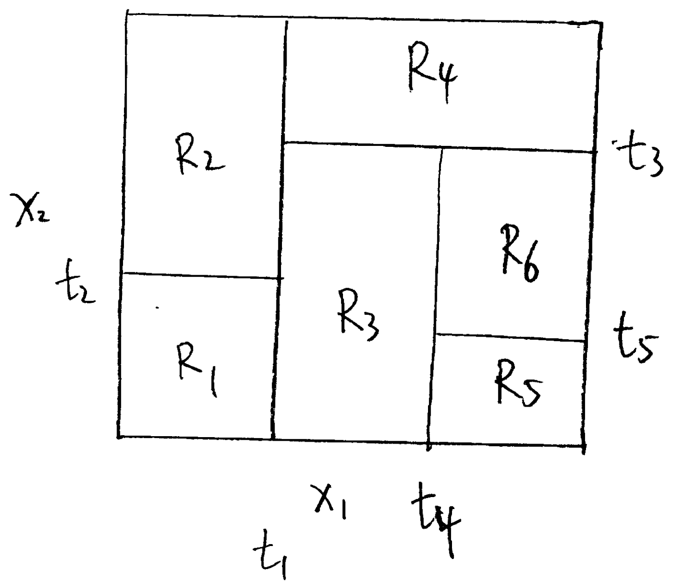
It is iterative because we estimate β one by one and update its value for every iteration.

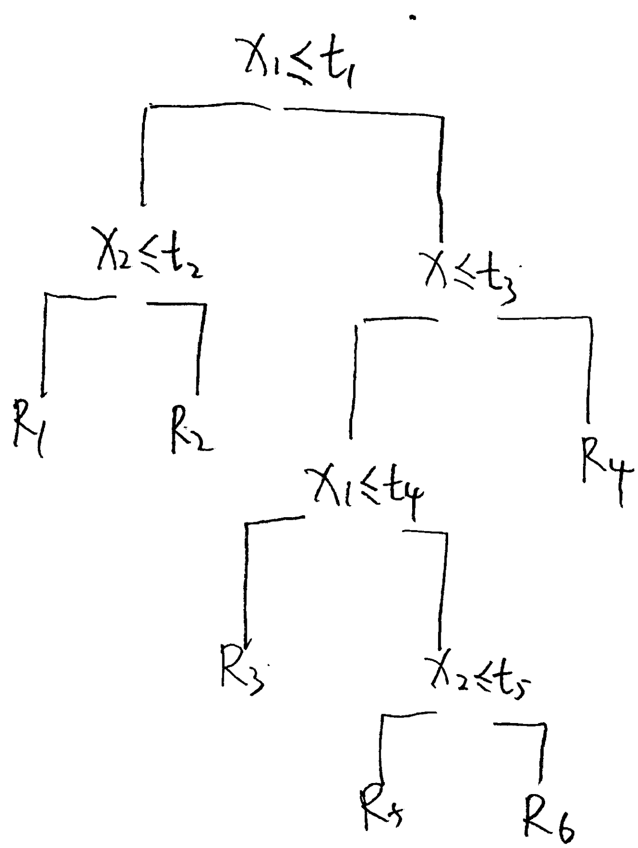
Besides, it can be seen as a weighted lease square estimator.

We reweighte matrix W every step, becawe $P = \frac{\exp^{\beta^{(m)}}}{1+\exp^{\beta^{(m)}}}$ is changing every step.

We update until convergence of β. With two properties, we call it IRWLS.

exercise 8.1





8.2 additive model takes the form $y_i = \beta_0 + f_1(X_{i1}) + f_2(X_{i2}) + \cdots + f_p(X_{ip}) + \varepsilon_i$

We can consider the example above. A single preditor has been cut more than once.

A single preditor X_j and its stump $\hat{f}_j(X_j) = \beta_0 + I(X_j < t_j)\beta_1$

1. In the beginning $\hat{f}(x) = 0$, $r_i = y_i$

2. (a) $\hat{f}^1(x) = \beta_{11} I(X_1 < t_1) + \beta_0$

(b) $\hat{f}(x) = \lambda \hat{f}^1(x)$

(c) $r_i = y_i - \lambda \hat{f}^1(x_i)$

continue for $1, 2, \cdots, B$ times

if a preditor has multiple stump, it can be seen as adding a branch to it, like the decision tree above. X_1 has 3 stumps t_1, t_3, t_5. Since all these stump functions solely depend on a single predictor. We can write multiple stump functions as one $\hat{f}_j(X_j)$ where

$\hat{f}_j(X_j) = \hat{f}_j(X_j)_1 + \hat{f}_j(X_j)_2$

And the final additive model is $f(X) = \sum_{j=1}^{p} f_j(X_j)$

where $f_j(X_j) = \frac{1}{\lambda} \hat{f}_j(X_j)$

exercise 10.1

$$\hat{\beta}_{t+1}, \hat{g}_{t+1} \leftarrow \underset{\beta, g}{\operatorname{argmin}} \sum_{i=1}^{n} \underbrace{e^{-y_i \hat{f}_t(x_i)}}_{w_i^{(t)}} e^{-y_i \beta g(x_i)}$$

$$= \underset{\beta, g}{\operatorname{argmin}} (e^{\beta}-e^{-\beta}) \underbrace{\sum_{i=1}^{n} w_i^{(t)} I(y_i \neq g(x_i))}_{E_g} + e^{-\beta} \underbrace{\sum_{i=1}^{n} w_i^{(t)}}_{W}$$

$$L = \underset{\beta, g}{\operatorname{argmin}} W\left((e^{\beta}-e^{-\beta})\frac{E_g}{W} + e^{-\beta}\right)$$

for given g, $\nabla_{\beta} L = E_g(e^{\beta}+e^{-\beta}) + (-We^{-\beta})$

set to 0

$$\frac{E_g}{W} = \frac{e^{-\beta}}{e^{\beta}+e^{-\beta}} = \frac{1}{e^{2\beta}+1}$$

$$\frac{W}{E_g} = e^{2\beta}+1$$

$$e^{2\beta} = \frac{W-E_g}{E_g} = \frac{1-E_g/W}{E_g/W}$$

$$\hat{\beta} = \frac{1}{2}\log\frac{1-E_g/W}{E_g/W}$$

10.2

$f^*(x) = \underset{f(x)}{\operatorname{argmin}}\ E_{Y|X}(e^{-Yf(x)})$ to find $f(x)$, take derivative and set it to 0

$$\frac{\partial}{\partial f} E_{Y|X}(e^{-Yf(x)}) = E_{Y|X}(-Ye^{-Yf(x)}) = 0$$

when $Y = \pm 1$, the above can be written as

$$-(-1)e^{-(-1)f(x)}Pr(Y=-1|x) - (1)e^{-(1)f(x)}Pr(Y=+1|x) = 0$$

$$e^{2f(x)}Pr(Y=-1|x) - Pr(Y=+1|x) = 0$$

$$e^{2f(x)} = \frac{Pr(Y=+1|x)}{Pr(Y=-1|x)}$$

$$f(x) = \frac{1}{2}\log\left(\frac{Pr(Y=+1|x)}{Pr(Y=-1|x)}\right)$$ one-half the log odds