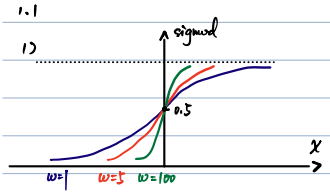
Written Problems

1.1



As the weight increases, the slope near x=0 become steeper. This indicates an intensive change in the sigmoid when X changes. Hence there is high possibility that the model would regard the noise as critic characteristics, leading to overfitting in many cases.

2) For logistic regression, $P(Y_i \mid X_i, w_0, w_1, \dots w_d) = \theta_i^{y_i}(1-\theta_i)^{1-y_i}$ where $\theta_i = (1+e^{-wx_i})^{-1}$

For $w \sim \mathcal{N}(0, \tau)$, $P(w_0, w_1, \dots w_d) = (2\pi)^{-\frac{d}{2}} \exp(-\frac{1}{2}\sum w_j^2)$

Hence the log-likelihood $\ln \prod_{i=1}^{n} P(Y_i \mid X_i, w)\, P(w)$

$= \ln \prod_{i=1}^{n} \left(\frac{1}{1+e^{-wx_i}}\right)^{y_i} \left(1-\frac{1}{1+e^{-wx_i}}\right)^{1-y_i} (2\pi)^{-\frac{d}{2}} \exp(-\frac{1}{2}\sum w_j^2)$

$= \sum_{i=1}^{n} y_i \ln\left(\frac{1}{1+e^{-wx_i}}\right) + (1-y_i)\ln\left(1-\frac{1}{1+e^{-wx_i}}\right) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\sum w_j^2$

$= -\frac{nd}{2}\ln 2\pi - \frac{n}{2}\sum w_j^2 + \sum_{i=1}^{n} y_i \ln\left(\frac{1}{1+e^{-wx_i}}\right) + (1-y_i)\ln\left(1-\frac{1}{1+e^{-wx_i}}\right)$

Drop those irrelavant with the weights. $\mathcal{J}(w) = -\frac{1}{n}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} y_i \ln\left(\frac{1}{1+e^{-wx_i}}\right) + (1-y_i)\ln\left(1-\frac{1}{1+e^{-wx_i}}\right)$

The gradient $\nabla_w \mathcal{J}(w) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{1+e^{-wx_i}}\right)x_i - w$

Hence we have the update rule:

(1) Initialize the vector w appropriately;

(2) Update w s.t. $w := w + \alpha \nabla_w \mathcal{J}(w)$, where $\nabla_w \mathcal{J}(w) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \frac{1}{1+e^{-wx_i}}\right)x_i - w$

(3) If $\|\nabla_w \mathcal{J}(w)\| \le \varepsilon$ or reaching the maximum # of iterations, stop.

Otherwise, execute (2) and (3) repeatedly.

M(C)AP addresses overfitting by introducing a penalty term as a prior, working as a regularization to the term of weights.

1.2

1) $P(y_c^i = 1 \mid x_i; w) = \frac{\exp(w_c^T x^i)}{\sum_{c'}\exp(w_{c'}^T x^i)} = \frac{\exp(w_c^T x^i - w_k^T x^i)\exp(w_k^T x_i)}{\sum_{c'}\exp(w_{c'}^T x^i - w_k^T x^i)\exp(w_k^T x_i)}$

$= \frac{\exp(w_c^T x^i - w_k^T x^i)}{\sum_{c'=1}^{k-1}\exp(w_{c'}^T x^i - w_k^T x^i) + \exp(w_k^T x_i - w_k^T x_i)} = \frac{\exp(w_c^T x^i - w_k^T x^i)}{\sum_{c'=1}^{k-1}\exp(w_{c'}^T x^i - w_k^T x^i) + 1}$

In the alternative expression, $w_c := w_c - w_k$? , $\forall c < k$

$P(y_c^i = 1 \mid x_i; w) = \frac{\exp(w_c^T x^i)}{\sum_{c'=1}^{k-1}\exp(w_{c'}^T x^i) + 1}$

$P(y_k^i = 1 \mid x_i; w) = 1 - \sum_{c'=1}^{k-1} P(y_{c'}^i = 1 \mid x_i; w) = \frac{1}{\sum_{c'=1}^{k-1}\exp(w_{c'}^T x^i) + 1}$

2) $\mathcal{L}(w) = \ln \prod_{j=1}^{n} P(y_c^j \mid x_c^j, w)$

$= \sum_{j=1}^{n} \ln \sum_{c'=1}^{k} y_{c'}^j P(y_{c'}^j \mid x_{c'}^j, w)$, with only the one with a true label $y_c^j = 1$.

$= \sum_{j=1}^{n} \ln \sum_{c'=1}^{k} y_{c'}^j \frac{\exp(w_{c'}^T x^j)}{\sum_{c'=1}^{k}\exp(w_{c'}^T x_j)}$, $\sum_{c'=1}^{k} y_{c'}^j = y_c^j$ hence the logarithm could move in.

$= \sum_{j=1}^{n}\sum_{c'=1}^{k} \ln y_{c'}^j \frac{\exp(w_{c'}^T x^j)}{\sum_{c'=1}^{k}\exp(w_{c'}^T x_j)}$

$= \sum_{j=1}^{n}\sum_{c'=1}^{k}\left[y_{c'}^j (w_{c'}^T x_j) - y_{c'}^j \ln\sum_{c'}\exp(w_{c'}^T x_j)\right]$

3) $\nabla_{w_c}\mathcal{L}_2(w) = \frac{\partial \sum_{c'=1}^{k} y_{c'}^j \ln(\sum_{c'}\exp(w_{c'}^T x_j))}{\partial \ln(\sum_{c'}\exp(w_{c'}^T x_j))} \cdot \frac{\partial \ln(\sum_{c'}\exp(w_{c'}^T x_j))}{\partial \sum_{c'}\exp(w_{c'}^T x_j)} \cdot \frac{\partial \sum_{c'}\exp(w_{c'}^T x_j)}{\partial w_c}$

$= y_c^j \cdot \frac{1}{\sum_{c'}\exp(w_{c'}^T x_j)} \cdot x_j \exp(w_c^T x_j)$

$= \frac{x_j \exp(w_c^T x_j)}{\sum_{c'}\exp(w_{c'}^T x_j)}$

4) $\frac{\partial \sum_j \sum_{c'} y_{c'}^j w_{c'}^T x^j}{\partial w_i} = \sum_j x^j y_c^j$, $y_c = 1$ only for the correct one.

Combining with $\nabla_{w_c}\mathcal{L}_2(w)$ yields $\nabla_{w_c}\mathcal{L}(w) = \sum_{j=1}^{n} x_j\left(y_c^j - P(y_c^j = 1 \mid x^j; w)\right)$

5) Hence we have the update rule:

(1) Initialize the vector w appropriately;

(2) Update w s.t. $w := w + \alpha \nabla_w \mathcal{L}(w)$, where $\nabla_w \mathcal{L}(w) = \sum_{j=1}^{n} x_j\left(y_c^j - P(y_c^j = 1 \mid x^j; w)\right)$

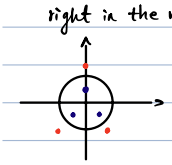
(3) If $\|\nabla_w \mathcal{L}(w)\| \le \varepsilon$ or reaching the maximum # of iterations, stop.

Otherwise, execute (2) and (3) repeatedly.

1.3

1) Recall $\|x\|_2 = \sqrt{x_1^2 + x_2^2}$, and in this case $\|x_1\|_2 = 1$, $\|x_2\|_2 = \sqrt{2}$

To find the largest possible geometric margin, it is intuitive to construct a circle right in the middle.



=> The appropriate SVM classifier could be $\|x\|_2 = \frac{1+\sqrt{2}}{2}$

2) In the $[x_1^2, x_2^2]$ space, we have $\begin{cases} \text{Class } -1: \{(0,1), (\frac{1}{2}, \frac{1}{2})\} \\ \text{Class } +1: \{(0,2), (1,1)\} \end{cases}$

$\min \frac{1}{2}\|w\|^2$ s.t. $\forall i\ y_i(w^T x_i + b) \ge 1$

By KKT, $\mathcal{L}(w, b, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m}\alpha_i[1 - y_i(w^T x_i + b)]$

$\frac{\partial \mathcal{L}}{\partial w} = 0$, $w = \sum_{i=1}^{m}\alpha_i y_i x_i$

$\frac{\partial \mathcal{L}}{\partial b} = 0$, $\sum_{i=1}^{m}\alpha_i y_i = 0$

$\alpha_i \ge 0$, $1 - y_i(w^T x + b) \le 0$

$\alpha_i(1 - y_i(w^T x + b)) = 0$

Hence the dual: $\max_{\alpha} \sum_i^m \alpha_i - \frac{1}{2}\sum_{i,j}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j$ s.t. $\forall i, \sum_{i=1}^{m}\alpha_i y_i = 0$, $\alpha_i \ge 0$

i.e. In this setting, $\max f(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ s.t. $-\alpha_1 - \alpha_2 + \alpha_3 + \alpha_4 = 0$

where $f(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}\alpha_1^2 - \frac{1}{4}\alpha_2^2 - 2\alpha_3^2 - \alpha_4^2$

$-\frac{1}{2}\alpha_1\alpha_2 + 2\alpha_1\alpha_3 + \alpha_1\alpha_4 + \alpha_2\alpha_3 + \alpha_2\alpha_4 - 2\alpha_3\alpha_4$

Thus $w = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$, $b = -3$

For $(-\frac{1}{2}, \sqrt{2})$, $2\times(\frac{1}{2})^2 + 2\times(\sqrt{2})^2 - 3 = \frac{3}{2} > 0$

=> It is predicted as class +1.

1.4

1) $\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\max(0, |y_i - w_i^T x_i| - \varepsilon)$

Converted into:

$\min \frac{1}{2}w^T w + C\sum_{i=1}^{n}\xi_i$

s.t. $|y_i - w_i^T x_i| - \varepsilon \le \xi_i$

$\xi_i \ge 0$

Further converted into:

$\min \frac{1}{2}w^T w + C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-)$

s.t. $y_i - w_i^T x_i - \varepsilon \le \xi_i^+$

$w_i^T x_i - y_i - \varepsilon \le \xi_i^-$

$\xi_i^+, \xi_i^- \ge 0$

2) Define the lagranging $\mathcal{L}(w, \xi^+, \xi^-, \alpha^+, \alpha^-, \mu^+, \mu^-)$

$= \frac{1}{2}w^T w + C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-) + \sum_i \alpha_i^+(y_i - w_i^T x_i - \varepsilon - \xi_i^+)$

$+\sum_i \alpha_i^-(w_i^T x_i - y_i - \varepsilon - \xi_i^-) - \sum_i \mu_i^+\xi_i^+ - \sum_i \mu_i^-\xi_i^-$

3) By KKT, we have conditions as follows:

Stationary: $\frac{\partial \mathcal{L}}{\partial w} = w - \sum_i \alpha_i^+ x_i + \sum_i \alpha_i^- x_i = 0$

$\frac{\partial \mathcal{L}}{\partial \xi_i^+} = C - \alpha_i^+ - \mu_i^+ = 0$ $\frac{\partial \mathcal{L}}{\partial \xi_i^-} = C - \alpha_i^- - \mu_i^+ = 0$

Feasibility: $\alpha_i^+, \alpha_i^- \ge 0$, $\xi^+, \xi^- \ge 0$, $\mu_i^+, \mu_i^- \ge 0$

Complementary: $\alpha_i^+(y_i - w_i^T x_i - \varepsilon - \xi_i^+) = 0$ $\alpha_i^-(w_i^T x_i - y_i - \varepsilon - \xi_i^-) = 0$

$\mu_i^+\xi_i^+ = 0$ $\mu_i^-\xi_i^- = 0$

Hence $\mathcal{L}(w) = \frac{1}{2}w^T w + C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-) + \sum_i \alpha_i^+(y_i - w_i^T x_i - \varepsilon - \xi_i^+)$

$+\sum_i \alpha_i^-(w_i^T x_i - y_i - \varepsilon - \xi_i^-) - \sum_i \mu_i^+\xi_i^+ - \sum_i \mu_i^-\xi_i^-$

$= \frac{1}{2}w^T w + C\sum_{i=1}^{n}(\xi_i^+ + \xi_i^-) + \sum_i(\alpha_i^+ - \alpha_i^-)y_i - \sum_i(\alpha_i^+ - \alpha_i^-)\left(\sum_j(\alpha_j^+ - \alpha_j^-)\right)$

$-\sum_i(\alpha_i^+ + \alpha_i^-)\varepsilon - \sum_i \alpha_i^+\xi_i^+ - \sum_i \alpha_i^-\xi_i^- - \sum_i \mu_i^+\xi_i^+ - \sum_i \mu_i^-\xi_i^-$

$= -\frac{1}{2}\sum_{i,j}(\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-)x_i^T x_j + \sum_i[(y_i - \varepsilon)x_i^+ - (y_i + \varepsilon)x_i^-]$

The dual could be written as

$\max -\frac{1}{2}\sum_{i,j}(\alpha_i^+ - \alpha_i^-)(\alpha_j^+ - \alpha_j^-)x_i^T x_j + \sum_i[(y_i - \varepsilon)x_i^+ - (y_i + \varepsilon)x_i^-]$

s.t. $0 \le \alpha_i^+ \le C$

$0 \le \alpha_i^- \le C$

4) Yes, obviously.

5) $(\alpha^+ = 0, \alpha^- \ne 0)$ or $(\alpha^+ \ne 0, \alpha^- = 0)$.

6) The prediction should be $\hat{y} = w^T x_i$, w as the trained weights.

7) Yes. replacing this $x_i^T x_j$ by $\phi(x_i, x_j)$.

8) $\varepsilon \downarrow$ => Width of boundary $\downarrow$ => Cost $\uparrow$

$\varepsilon \uparrow$ => Width of boundary $\uparrow$ => Lose some feature but cost $\downarrow$

9) $C \uparrow$ => More penalty on outliers instead of regularization => More inclusive

$C \downarrow$ => More penalty on regularization => The hyperplane is flatter.

1.5

$H(D) = -\frac{6}{15}\log_2\frac{6}{15} - \frac{9}{15}\log_2\frac{9}{15} = 0.971$

Age: $H(D|\text{Age}=\text{young}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$

$H(D|\text{Age}=\text{middle}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$

$H(D|\text{Age}=\text{old}) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.722$

$g(D,A) = H(D) - H(D|A) = 0.971 - 0.971\times\frac{1}{3} - 0.971\times\frac{1}{3} - 0.722\times\frac{1}{3} = 0.083$

$H_A(D) = -\frac{5}{15}\log_2\frac{5}{15} - \frac{5}{15}\log_2\frac{5}{15} - \frac{5}{15}\log_2\frac{5}{15} = 1.585$

$g_R(D,A) = \frac{g(D,A)}{H_A(D)} = \frac{0.083}{1.585} = 0.052$

Work: $H(D|\text{Work}=\text{yes}) = -\log_2 1 - 0 = 0$

$H(D|\text{Work}=\text{no}) = -\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10} = 0.971$

$g(D,A) = H(D) - H(D|A) = 0.971 - \frac{2}{3}\times 0.971 = 0.324$

$H_A(D) = -\frac{10}{15}\log_2\frac{10}{15} - \frac{5}{15}\log_2\frac{5}{15} = 0.918$

$g_R(D,A) = \frac{g(D,A)}{H_A(D)} = 0.353$

Similarly, for "House" $g_R(D,A) = 0.433$; for "Credit" $g_R(D,A) = 0.232$.

Hence we first divide it using "House". The ones with "House" labelled as yes all, so we continue to divide the ones with "House" labelled as no.

In the next stage, we similarly have $H(D) = 0.918$

The "Age" $g_R(D,A) = 0.151$, the "Work" $g_R(D,A) = 1$, and the "Credit" $g_R(D,A) = 0.341$

Hence we next divide it using "Work".

As a result,