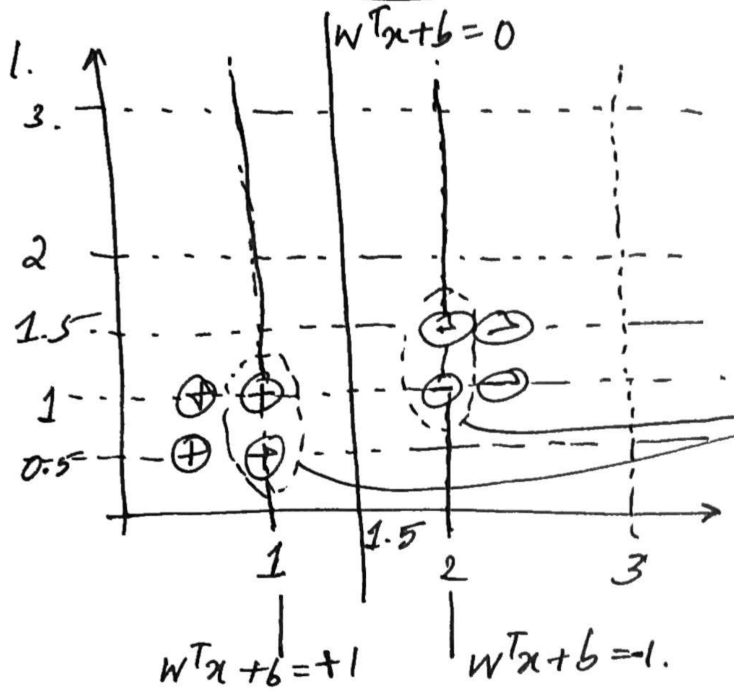
MACHINE LEARNING – WA-3

1.



a. The points $(1, 0.5)$ and $(1,1)$ are the support vectors for $w^Tx+b = +1$, and $(2,1)$ and $(2, 1.5)$ are the support vectors for $w^Tx+b=-1$.

b. We can see from the figure $w^Tx+b=0$ passes through 1.5. Also we need the distance to the support vectors $=1$ and the distance of all other points ≥ 1 (ie) $y_i(w^Tx_i+b) \geq 1$

$\Rightarrow$ Let the decision boundary be of the form

$$W_1 x + W_2 y + W_3 = 0$$

We know $W_2 = 0$ (from the nature of our decision boundary $x=1.5$) and ~~at~~ $(1.5, 0)$ lies on the boundary

$$\Rightarrow W_1 x - 1.5W_1 = 0.$$

Distance of $(1,1)$ from $w^Tx+b=+1$ = distance of $(2,1)$ from $w^Tx+b=-1$

$\Rightarrow$ ~~$w_1 + 5w_1$~~ $y_i(w_1 x_i - 1.5w_1) = 1$ for $(1,1)$ and $(2,1)$

$\Rightarrow +1(w_1 1 - 1.5w_1) = -1(2w_1 - 1.5w_1) = -0.5w_1 = 1$

$\Rightarrow w_1 = \frac{-1}{0.5} = \underline{\underline{-2}}$

$\Rightarrow$ The decision boundary is $-2x + 3 = 0$

$\underline{\underline{W = [-2, 0]^T \quad b = +3}}$

2. a. Consider the optimal solution (W^*, b^*, ξ_i^*).

Assume that $\forall i \; \xi_i \leq 0$.

$\Rightarrow 1 - \xi_i > 1$

$\Rightarrow$ The decision boundaries $y_i(W^T x_i + b) \geq 1 - \xi_i > 1$

$\Rightarrow$ For all elements (x_i, y_i) this is equivalent to having $\xi_i = 0$ as $y_i(W^T x_i + b) \geq 1$.

Thus any value of ξ_i in the solution, where $\xi_i < 0$ is equivalent to $\xi_i = 0$, as it only increases the distance of the point from the original boundary. Also as the objective is $W^T W + \sum_i \lambda |\xi_i|^2$, the sign of ξ_i does not affect the optimization.

Thus all $\xi_i < 0$, play no role in the optimization, and can be replaced by $\xi_i = 0$.

$\Rightarrow$ Even without the final constraint all ξ_i in our solution would be greater / equal to 0.

b. $f(W, \xi) = W^TW + \lambda \sum_i \xi_i^2.$

$$1-\xi_i \le y_i(W^Tx_i+b) \Rightarrow 1-\xi_i - y_i(W^Tx_i+b) \le 0$$

$$\mathcal{L}(W, b, \xi, \alpha) = W^TW + \lambda \sum_i \xi_i^2 + \sum_i \alpha_i\left(1-\xi_i - y_i(W^Tx_i+b)\right)$$

c. Primal: $\min_{W, b, \xi} \max_{\alpha} W^TW + \lambda \sum_i \xi_i^2 + \sum_i \alpha_i\left(1-\xi_i - y_i(W^Tx_i+b)\right)$

Dual $\max_{\alpha} \min_{W, b, \xi} W^TW + \lambda \sum_i \xi_i^2 + \sum_i \alpha_i\left(1-\xi_i - y_i(W^Tx_i+b)\right)$

Similar to the previous case $W = \frac{1}{2}\sum_i \alpha_i y_i x_i$ and $\sum_i \alpha_i y_i = 0$

Consider the optimization $\min_{W, b, \xi} W^TW + \lambda \sum_i \xi_i$

subject to $y_i(W^Tx_i+b) \ge 1-\xi_i \quad \forall i.$

The constraint can be rewritten as $\xi_i \ge 1-y_i(W^Tx_i+b)$

$$\Rightarrow \xi_i = \underbrace{\max(0, 1-y_i(W^Tx_i+b))}_{\text{Hinge Loss.}}$$

$\Rightarrow$ Optimization $\min_{W, b} \|W\|^2 + \lambda \sum_{i=1}^{N} \left(\max(0, 1-y_i(W^Tx_i+b))\right)^2$

The L_2 SVM is similar to using an L_2 regularization on ξ whereas the standard SVM with a hinge loss can be viewed as being similar to L_1 regularization on ξ.

$$\mathcal{L}(w,b,\xi,\alpha) = w^T w + \lambda \sum_i \xi_i^2 + \cancel{2}\sum_i \alpha_i\left(1 - \xi_i - y_i(w^T x_i + b)\right)$$

$$\frac{\partial L}{\partial w} : 2w - \sum \alpha_i y_i x_i = 0 \Rightarrow w = \frac{1}{2}\sum_i \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum \alpha_i y_i = 0$$

Substituting for w:

$$\frac{1}{4}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j + \lambda \sum_i \xi_i^2 + \sum_i \alpha_i - \sum_i \alpha_i \xi_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j \underbrace{- b \sum_i \alpha_i y_i}_{0}$$

$$\Rightarrow \mathcal{L}(w,b,\xi,\alpha) = \lambda \sum_i \xi_i^2 + \sum_i \alpha_i - \frac{1}{4}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j$$

$$\cancel{\text{Dual}: \max_\alpha \lambda \sum_i}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \Rightarrow 2\lambda \sum_i \xi_i - \sum_i \alpha_i = 0$$

$$\Rightarrow \lambda = \frac{1}{2} \frac{\sum_i \alpha_i}{\sum_i \xi_i}$$

$$\Rightarrow \mathcal{L}(w,b,\xi,\alpha) = \cancel{\frac{1}{2}\sum_i \alpha_i \xi_j}\ \frac{1}{2}\sum_i \sum_j \alpha_i \xi_j + \sum_i \alpha_i - \frac{1}{4}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j$$

Dual: $\max_\alpha \frac{1}{2}\sum_i \sum_j \alpha_i \xi_j + \sum_i \alpha_i - \frac{1}{4}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j$

However the biggest byproduct of this formulation is the sparsity. The new L_2-Hinge loss is less sparse in comparison to the previous L_1-Hinge Loss. As a result of lesser sparsity of ξ_i in the L_2 SVM, there will be more support vectors in L_2-SVM than the L_1 version. Therefore the L_2-version is more prone to outliers.

3. a. $P(A=1/y=0) = 1/3$

$P(A=1/y=1) = 2/3$

$P(B=1/y=0) = 2/3$

$P(B=1/y=1) = 2/3$

$P(C=1/y=0) = 2/3$

$P(C=1/y=1) = 1/3$

$P(y=1) = 1/2$

b. $$P(y=1/A=1, B=0, C=0) = \frac{P(A=1,B=0,C=0/y=1) \cdot P(y=1)}{P(A=1,B=0,C=0/y=1) + P(A=1,B=0,C=0/y=0)}$$

$$= \frac{2/3 \times 1/3 \times 2/3 \times 1/2}{2/3 \times 1/3 \times 2/3 \times 1/2 + 1/3 \times 1/3 \times 1/3 \times 1/2} = \frac{4/54}{4/54 + 1/54} = \frac{4}{5}$$

c. As discussed in class conditional independence and independence are not equal. This can be seen from the Rock-Paper-Scissors game, where the choice of state is independent but given the result of the game, each input are not ~~mutually~~ conditionally independent. So the Naive Bayes assumption need not be valid if A, B, C are independent random variables.

4. We predict the output to be class 1 if $P(y=1|x) \geq P(y=0|x)$

$$\Rightarrow \frac{P(y=1|x)}{P(y=0|x)} \geq 1 \quad \Rightarrow \log\left(\frac{P(y=1|x)}{P(y=0|x)}\right) \geq 0$$

$$\Rightarrow \log\left(\frac{P(x|y=1).\,P(y=1)}{P(x|y=0).\,P(y=0)}\right) \geq 0$$

Because each feature x_i is a binary variable we can write

$P(x_i|y=1/0)$ as $\theta_{i1/0}^{x_i}\,(1-\theta_{i1/0}^{(1-x_i)})$

$$\Rightarrow \log\left(\frac{\theta_{11}^{x_1}.(1-\theta_{11})^{1-x_1}.\ \theta_{21}^{x_2}(1-\theta_{21})^{1-x_2}\ldots\theta_{d1}^{x_d}(1-\theta_{d1})^{1-x_d}.\,P(y=1)}{\theta_{10}^{x_1}(1-\theta_{10})^{1-x_1}\ \theta_{20}^{x_2}(1-\theta_{20})^{1-x_2}\ldots\theta_{d0}^{x_d}(1-\theta_{d0})^{1-x_d}\,P(y=0)}\right) \geq 0$$

$$\Rightarrow x_1\log\left(\frac{\theta_{11}}{\theta_{10}}\right) + (1-x_1)\log\left(\frac{1-\theta_{11}}{1-\theta_{10}}\right) + x_2\log\left(\frac{\theta_{21}}{\theta_{20}}\right) + (1-x_2)\log\left(\frac{1-\theta_{21}}{1-\theta_{20}}\right)$$

$$+\cdots\ x_d\log\left(\frac{\theta_{d1}}{\theta_{d0}}\right) + (1-x_d)\log\left(\frac{1-\theta_{d1}}{1-\theta_{d0}}\right) + \log\left(\frac{P(y=1)}{P(y=0)}\right) \geq 0$$

$$\Rightarrow x_1\left(\log\left(\frac{\theta_{11}}{\theta_{10}}\right) - \log\left(\frac{1-\theta_{11}}{1-\theta_{10}}\right)\right) + x_2\left(\log\left(\frac{\theta_{21}}{\theta_{20}}\right) - \log\left(\frac{1-\theta_{21}}{1-\theta_{20}}\right)\right) + \cdots$$

$$\cdots\ x_d\left(\log\left(\frac{\theta_{d1}}{\theta_{d0}}\right) - \log\left(\frac{1-\theta_{d1}}{1-\theta_{d0}}\right)\right) +$$

$$\left(\log\left(\frac{1-\theta_{11}}{1-\theta_{10}}\right) + \log\left(\frac{1-\theta_{21}}{1-\theta_{20}}\right) + \cdots \log\left(\frac{1-\theta_{d1}}{1-\theta_{d0}}\right) + \log\left(\frac{P(y=1)}{P(y=0)}\right)\right) \geq 0$$

$$\Rightarrow w_0 = \log\left(\frac{P(y=1)}{P(y=0)}\right) + \log\left(\frac{1-\theta_{11}}{1-\theta_{10}}\right) + \log\left(\frac{1-\theta_{21}}{1-\theta_{20}}\right) + \cdots \log\left(\frac{1-\theta_{d1}}{1-\theta_{d0}}\right)$$

$$w_i = \log\left(\frac{\theta_{i1}}{\theta_{i0}}\right) - \log\left(\frac{1-\theta_{i1}}{1-\theta_{i0}}\right) \quad \forall\, i = 1 \cdots d$$