EE566 ADAPTATION AND LEARNING **Instructor: Ali H. Sayed**
Homework Assignment #4 **Due: May 26, 2025**

---

1) (**Chapter 55**) Consider $M-$dimensional feature vectors with independent Boolean entries, $\boldsymbol{h} \in \{0,1\}^M$. Each feature vector belongs to one of two classes, denoted by $\boldsymbol{\gamma} \in \{\pm 1\}$ with $\mathbb{P}(\boldsymbol{\gamma} = +1) = \pi > 0$. Assume the individual entries of $\boldsymbol{h}$ are distributed according to the probabilities:

$$\mathbb{P}(\boldsymbol{h}(m) = 1|\boldsymbol{\gamma} = +1) = \theta_{+m} > 0, \qquad \mathbb{P}(\boldsymbol{h}(m) = 1|\boldsymbol{\gamma} = -1) = \theta_{-m} > 0$$

Under the independence assumption, the Bayes classifier reduces to the naïve Bayes construction. We wish to relate the latter in this problem to the logistic regression form. Thus, given an arbitrary feature vector h, the Bayes classifier determines its class by seeking the label that solves:

$$\gamma^{\bullet}(h) = \underset{\gamma \in \{\pm 1\}}{\text{argmax}} \ \Big\{ \pi_{\gamma} \, \mathbb{P}(\boldsymbol{h} = h|\boldsymbol{\gamma} = \gamma) \Big\}$$

where $\pi_{+1} = \pi$ and $\pi_{-1} = 1 - \pi$. Show that, in this case, the posterior probability of the label given the feature vector takes the following form in terms of the logistic function:

$$\mathbb{P}(\boldsymbol{\gamma} = +1|\boldsymbol{h} = h) \ = \ \sigma(h^{\mathsf{T}} w - \theta)$$

where $\sigma(z) = 1/(1 + e^{-z})$ and

$$\begin{aligned} -\theta \ &\triangleq \ \ln\Big(\frac{\pi}{1-\pi}\Big) + \sum_{m=1}^{M} \ln\Big(\frac{1-\theta_{+m}}{1-\theta_{-m}}\Big) \\ w_m \ &\triangleq \ \ln\Big(\frac{\theta_{+m}(1-\theta_{-m})}{\theta_{-m}(1-\theta_{+m})}\Big), \quad m = 1, 2, \ldots, M \end{aligned}$$

Solution: Using Bayes rule we have that

$$\begin{aligned} \mathbb{P}(\boldsymbol{\gamma} = +1|\boldsymbol{h} = h) \ &= \ \frac{\mathbb{P}(\boldsymbol{\gamma} = +1)\mathbb{P}(\boldsymbol{h} = h|\boldsymbol{\gamma} = +1)}{\mathbb{P}(\boldsymbol{h} = h)} \\ \mathbb{P}(\boldsymbol{\gamma} = -1|\boldsymbol{h} = h) \ &= \ \frac{\mathbb{P}(\boldsymbol{\gamma} = -1)\mathbb{P}(\boldsymbol{h} = h|\boldsymbol{\gamma} = -1)}{\mathbb{P}(\boldsymbol{h} = h)} \end{aligned}$$

where

$$\begin{aligned} \mathbb{P}(\boldsymbol{\gamma} = +1) \ &= \ \pi \\ \mathbb{P}(\boldsymbol{\gamma} = -1) \ &= \ 1 - \pi \\ \mathbb{P}(\boldsymbol{h} = h|\boldsymbol{\gamma} = +1) \ &= \ \prod_{m=1}^{M} \theta_{+m}^{h(m)} (1 - \theta_{+m})^{1-h(m)} \\ \mathbb{P}(\boldsymbol{h} = h|\boldsymbol{\gamma} = -1) \ &= \ \prod_{m=1}^{M} \theta_{-m}^{h(m)} (1 - \theta_{-m})^{1-h(m)} \end{aligned}$$

Using the fact that

$$\mathbb{P}(\boldsymbol{\gamma} = +1|\boldsymbol{h} = h) + \mathbb{P}(\boldsymbol{\gamma} = -1|\boldsymbol{h} = h) = 1$$

we can solve for $\mathbb{P}(\boldsymbol{h} = h)$ and find that

$$\mathbb{P}(\boldsymbol{h} = h) = \pi \prod_{m=1}^{M} \theta_{+m}^{h(m)} (1 - \theta_{+m})^{1-h(m)} \ + \ (1 - \pi) \prod_{m=1}^{M} \theta_{-m}^{h(m)} (1 - \theta_{-m})^{1-h(m)}$$

It follows that

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{\gamma}=+1|\boldsymbol{h}=h)\\
&=\frac{\pi\prod_{m=1}^{M}\theta_{+m}^{h(m)}(1-\theta_{+m})^{1-h(m)}}{\pi\prod_{m=1}^{M}\theta_{+m}^{h(m)}(1-\theta_{+m})^{1-h(m)}\ +\ (1-\pi)\prod_{m=1}^{M}\theta_{-m}^{h(m)}(1-\theta_{-m})^{1-h(m)}}\\
&=\frac{1}{1+\exp\Big(\log\Big(\frac{\pi}{1-\pi}\Pi_{m=1}^{M}\Big(\frac{\theta_{-m}}{\theta_{+m}}\Big)^{h(m)}\Big(\frac{1-\theta_{-m}}{1-\theta_{+m}}\Big)^{1-h(m)}\Big)\Big)}
\end{aligned}$$

Let

$$\begin{aligned}
z&=-\log\Big(\frac{\pi}{1-\pi}\prod_{m=1}^{M}\Big(\frac{\theta_{-m}}{\theta_{+m}}\Big)^{h(m)}\Big(\frac{1-\theta_{-m}}{1-\theta_{+m}}\Big)^{1-h(m)}\Big)\\
&=\log\Big(\frac{\pi}{1-\pi}\Big)+\sum_{m=1}^{M}\log\left(\Big(\frac{\theta_{+m}}{\theta_{-m}}\Big)^{h(m)}\Big(\frac{1-\theta_{+m}}{1-\theta_{-m}}\Big)^{1-h(m)}\right)\\
&=\log\Big(\frac{\pi}{1-\pi}\Big)+\sum_{m=1}^{M}\log\frac{1-\theta_{+m}}{1-\theta_{-m}}+\sum_{m=1}^{M}h(m)\log\frac{\theta_{+m}(1-\theta_{-m})}{\theta_{-m}(1-\theta_{+m})}
\end{aligned}$$

Thus, we take

$$\begin{aligned}
-\theta&\ \triangleq\ \ln\Big(\frac{\pi}{1-\pi}\Big)+\sum_{m=1}^{M}\ln\Big(\frac{1-\theta_{+m}}{1-\theta_{-m}}\Big)\\
w_m&\ \triangleq\ \ln\Big(\frac{\theta_{+m}(1-\theta_{-m})}{\theta_{-m}(1-\theta_{+m})}\Big),\quad m=1,2,\ldots,M
\end{aligned}$$

□

2) (**Chapter 57**) Consider the rank-1 approximation problem

$$x^o\ =\ \underset{x\in\mathbb{R}^{N\times 1}}{\operatorname{argmin}}\ \frac{1}{4}\|H-xx^{\mathsf{T}}\|_{\mathrm{F}}^2$$

where H is $N\times N$ symmetric and positive definite. If λ_1 is the largest eigenvalue of H with unit-norm eigenvector u_1, i.e., $Hu_1=\lambda_1 u_1$, then we know by inspection that one solution is $x^o=\sqrt{\lambda_1}u_1$. In this problem we wish to examine a gradient-descent recursion for learning a solution. Let λ_2 denote the second largest eigenvalue of H and assume $\lambda_1>\lambda_2$.

(a) Is the cost function convex over x? Describe all stationary points of the optimization problem.
(b) Show that the gradient-descent recursion takes the form

$$x_n=x_{n-1}-\mu_n(x_{n-1}x_{n-1}^{\mathsf{T}}-H)x_{n-1}$$

(c) Select $\mu_n=\mu/(1+\mu\|x_{n-1}\|^2)$ where $\mu>0$. Show that the recursion of part (b) reduces to

$$x_n=\frac{1}{1+\mu\|x_{n-1}\|^2}(I+\mu H)x_{n-1}$$

(d) Argue that after N iterations, the direction of x_N converges to

$$\frac{x_N}{\|x_N\|}=\frac{(I_N+\mu H)^{N+1}x_{-1}}{\|I_N+\mu H)^{N+1}x_{-1}\|}$$

Conclude that x_N converges toward the direction of u_1 and argue that the algorithm essentially

reduces to

$$x_n = \left(\frac{1+\mu\lambda_1}{1+\mu\|x_{n-1}\|^2}\right)x_{n-1}, \quad \text{large } n$$

Solution:

(a) Let $P(x) = \frac{1}{4}\|H - xx^{\mathsf{T}}\|_{\mathrm{F}}^2$ denote the cost function. This is not convex. Note in particular that $\sqrt{\lambda_1}u_1$ and $-\sqrt{\lambda_1}u_1$ are 2 different global minimizers.

We introduce the eigendecomposition of H:

$$H = \sum_{n=1}^{N} \lambda_n u_n u_n^{\mathsf{T}}$$

where the $\{u_n\}$ are orthonormal. The gradient vector and the Hessian matrix of the cost function are given by

$$\begin{aligned}
\nabla_{x^{\mathsf{T}}} P(x) &= \nabla_{x^{\mathsf{T}}} \left\{\frac{1}{4}\Big(\|x\|^4 - 2x^{\mathsf{T}}Hx + \|H\|_{\mathrm{F}}^2\Big)\right\} = xx^{\mathsf{T}}x - Hx \\
\nabla_x^2 P(x) &= \|x\|^2 I_N + 2xx^{\mathsf{T}} - H
\end{aligned}$$

Therefore, all stationary points should satisfy $Hx^o = \|x^o\|^2 x^o$. These include $x^o = 0$ as well as $x^o = \pm\sqrt{\lambda_n}u_n$, where the (λ_n, u_n) denote eigenvalue-eigenvector pairs for H. We know that the global minima are given by $\pm\sqrt{\lambda_1}u_1$. For the remaining stationary points $x^o \notin \{\pm\sqrt{\lambda_1}u_1\}$, the Hessian matrix satisfies

$$\begin{aligned}
u_1^{\mathsf{T}}\nabla_x^2 P(x^o)u_1 &= u_1^{\mathsf{T}}\Big((x^o)^{\mathsf{T}}x^o I_N + 2x^o(x^o)^{\mathsf{T}} - H\Big)u_1 \\
&= u_1^{\mathsf{T}}(x^o)^{\mathsf{T}}x^o u_1 - u_1^{\mathsf{T}}Hu_1, \quad \text{since } u_1^{\mathsf{T}}x^o = 0 \\
&= \|u_1\|^2 \times \|x^o\|^2 - u_1^{\mathsf{T}}Hu_1 \\
&\leq \lambda_2 - \lambda_1, \quad \text{since } \|u_1\|^2 = 1 \text{ and } \|x^o\|^2 \leq \lambda_2 \\
&< 0
\end{aligned}$$

This means that the Hessian matrix has a negative curvature along the u_1 direction. Thus, the other stationary points are saddle points.

(b) The gradient vector relative to x is given by

$$\nabla_{x^{\mathsf{T}}} P(x) = xx^{\mathsf{T}}x - Hx$$

and, hence,

$$x_n = x_{n-1} - \mu_n(x_{n-1}x_{n-1}^{\mathsf{T}} - H)x_{n-1}$$

(c) Using the selection for μ_n we get

$$\begin{aligned}
x_n &= x_{n-1} - \frac{\mu}{1+\mu\|x_{n-1}\|^2}(x_{n-1}x_{n-1}^{\mathsf{T}} - H)x_{n-1} \\
&= \Big(1 - \frac{\mu}{1+\mu\|x_{n-1}\|^2}\|x_{n-1}\|^2\Big)x_{n-1} + \frac{\mu}{1+\mu\|x_{n-1}\|^2}Hx_{n-1} \\
&= \frac{1}{1+\mu\|x_{n-1}\|^2}\Big(x_{n-1} + \mu Hx_{n-1}\Big) \\
&= \frac{1}{1+\mu\|x_{n-1}\|^2}\Big(I_N + \mu H\Big)x_{n-1}
\end{aligned}$$

(d) Iterating we get

$$x_N = (I + \mu H)^{N+1} x_{-1} \prod_{n=0}^{N} \frac{1}{1 + \mu \|x_{n-1}\|^2}$$

and, hence,

$$\frac{x_N}{\|x_N\|} = \frac{(I_N + \mu H)^{N+1} x_{-1}}{\|I_N + \mu H)^{N+1} x_{-1}\|}$$

The numerator has a form similar to the power iteration: Running $r_n = A r_{n-1}$, which leads to $r_n = A^{n+1} r_{-1}$, converges towards an eigenvector for A corresponding to its largest eigenvalue. Thus, x_N converges towards the direction of the eigenvector of $I_N + \mu H$ corresponding to the largest eigenvalue at $1 + \mu\lambda_1$; this eigenvector is parallel to u_1. In this case,

$$(I + \mu H) x_{n-1} \to (1 + \mu\lambda_1) x_{n-1}$$

and the recursion reduces to

$$x_n = \frac{(1 + \mu\lambda_1)}{1 + \mu \|x_{n-1}\|^2} x_{n-1}$$

□

3) (**Chapter 59**) Consider feature vectors $h \in \mathbb{R}^M$ and a collection of K classifiers (or experts) denoted by $\{E_1(h), E_2(h), \ldots, E_K(h)\}$. Each feature vector $h \in \mathbb{R}^M$ can belong to one of R classes denoted by $r = 1, 2, \ldots, R$. Introduce an $R \times K$ matrix $\mathcal{E}$, which summarizes the opinion of the experts about the class of h. Each row of index r corresponds to one of the labels, and each column of index c corresponds to one of the classifiers or experts. The entry $\mathcal{E}_{rc}$ indicates the level of confidence that expert E_c has about feature h belonging to class r. For illustration purposes, we exhibit a matrix $\mathcal{E}$ corresponding to $R = 4$ labels and $K = 5$ experts:

$$\mathcal{E} = \left[\begin{array}{c|ccccc} \text{labels} & E_1 & E_2 & E_3 & E_4 & E_5 \\ \hline r = 1 & \mathcal{E}_{11} & \mathcal{E}_{12} & \mathcal{E}_{13} & \mathcal{E}_{14} & \mathcal{E}_{15} \\ r = 2 & \mathcal{E}_{21} & \mathcal{E}_{22} & \mathcal{E}_{23} & \mathcal{E}_{24} & \mathcal{E}_{25} \\ r = 3 & \mathcal{E}_{31} & \mathcal{E}_{32} & \mathcal{E}_{33} & \mathcal{E}_{34} & \mathcal{E}_{35} \\ r = 4 & \mathcal{E}_{41} & \mathcal{E}_{42} & \mathcal{E}_{43} & \mathcal{E}_{44} & \mathcal{E}_{45} \end{array}\right]$$

There are many ways by which the information in $\mathcal{E}$ can be fused together to arrive at a recommendation for the label $r^\star$ for h.

(a) The majority of votes approach operates as follows. Each classifier makes its own decision, and subsequently the label $r^\star$ that receives the most votes is selected. Argue that this construction amounts to carrying out the following calculations. The matrix $\mathcal{E}$ is transformed into a new $R \times K$ matrix $\mathcal{D}$. Each column of $\mathcal{D}$ is a basis vector with a unit entry at the location corresponding to the largest confidence level for that classifier, and 0 elsewhere. If more that one label corresponds to the highest confidence level, we select at random one of the entries. Subsequently, we set

$$r^\star = \underset{1 \le r \le R}{\text{argmax}} \ \|\mathcal{D}\|_\infty$$

(b) A second approach is to combine the confidence levels of all classifiers for each label and then select $r^\star$ as the label corresponding to the highest aggregate score. Argue that this amounts to computing

$$r^\star = \underset{1 \le r \le R}{\text{argmax}} \ \|\mathcal{E}\|_\infty$$

(c) Sometimes, a normalization step is included in order to ensure that the confidence levels across all experts are comparable to each other. In this case, each entry $\mathcal{E}_{rc}$ is replaced by the softmax

value

$$\mathcal{E}'_{rc} \leftarrow \frac{e^{\mathcal{E}_{rc}}}{\sum_{r=1}^{R} e^{\mathcal{E}_{r'c}}}$$

Explain that the choice of $r^\star$ now results from

$$r^\star = \underset{1\leq r\leq R}{\text{argmax}} \prod_{c=1}^{C} \mathcal{E}'_{rc}$$

(d) A third approach corresponds to using $\mathcal{E}$ to perform a feature transformation. Specifically, each feature h is replaced by a new feature vector $h' = \text{col}\{\mathcal{E}\}$ by stacking the columns of $\mathcal{E}$ on top of each other and using the features $\{h'\}$ to train a new classifier. What is the dimension of the transformed feature space?

Solution: The reader may refer to Bicego and Loog (2016).

(a) Each classifier c places a 1 at the location corresponding to the largest confidence level in its column and 0's elsewhere. The row of the largest confidence level for classifier c is found through the calculation:

$$r' = \underset{1\leq r\leq R}{\text{argmax}}\ \mathcal{E}_{rc}$$

Then the matrix $\mathcal{D}$ would look like the following example, with the 1's placed at the locations (r', c) for each expert:

$$\mathcal{D} = \left[\begin{array}{c|ccccc} \text{labels} & E_1 & E_2 & E_3 & E_4 & E_5 \\ \hline r=1 & 1 & 1 & 0 & 1 & 0 \\ r=2 & 0 & 0 & 1 & 0 & 0 \\ r=3 & 0 & 0 & 0 & 0 & 1 \\ r=4 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

In this example, the label that receives the most votes is $r = 1$.

(b) By adding the confidence scores across each row, we determine an aggregate value for the confidence levels of all experts in that particular label. By selecting the row with the largest aggregate sum, we are in effect determining the row that corresponds to the $\infty-$norm of $\mathcal{E}$ so that

$$r^\star = \underset{1\leq r\leq R}{\text{argmax}}\ \|\mathcal{E}\|_\infty$$

(c) Note that

$$\begin{aligned}
\underset{1\le r\le R}{\text{argmax}} \prod_{c=1}^{C} \mathcal{E}'_{rc} &= \underset{1\le r\le R}{\text{argmax}} \ \ln\left\{\prod_{c=1}^{C} \mathcal{E}'_{rc}\right\} \\
&= \underset{1\le r\le R}{\text{argmax}} \sum_{c=1}^{C} \ln \mathcal{E}'_{rc} \\
&= \underset{1\le r\le R}{\text{argmax}} \sum_{c=1}^{C} \ln \frac{e^{\mathcal{E}_{rc}}}{\sum_{r=1}^{R} \mathcal{E}_{rc}} \\
&= \underset{1\le r\le R}{\text{argmax}} \sum_{c=1}^{C} \mathcal{E}_{rc} \\
&= \underset{1\le r\le R}{\text{argmax}} \ \|\mathcal{E}\|_{\infty} \\
&= r^{\star}
\end{aligned}$$

(d) The size of h' is RK while the dimension of the original feature space is M.

□

4) (**Chapter 60**) Consider a linearly separable dataset $\{\gamma(n), h_n\}$ of size N where the labels $\gamma(n) \in \{-1, +1\}$ and the feature vectors h_n are $M-$dimensional. Linear separability guarantees the existence of a separation hyperplane $\{w^{\star}, \theta^{\star}\}$ such that $\gamma(n)(h_n^{\mathsf{T}} w^{\star} - \theta^{\star}) > 0$ for $n = 0, 1, \ldots, N-1$. For convenience, we extend the feature vectors and the weight vector using construction (60.20) so that the separability condition ensures $\gamma(n) h_n^{\mathsf{T}} w^{\star} > 0$. Without loss of generality, we normalize $w^{\star}$ to $\|w^{\star}\| = 1$ and rescale the feature vectors to satisfy $\|h_n\| \le 1$. In other words, the $w^{\star}$ lies on the unit sphere and the h_n lies within the sphere. We apply the perceptron algorithm with step size parameter $\mu = 1$ to determine a separating hyperplane, starting from the zero iterate. Let w denote the current iterate value and assume the feature vector h is misclassified. The iterate w is then updated to $w_u \leftarrow w + \gamma h$. Verify that

(a) The margin value is given by $m(w^{\star}) = \min_{0\le n\le N-1} |h_n^{\mathsf{T}} w^{\star}|$.
(b) The inner product of the iterate with $w^{\star}$ satisfies $w_u^{\mathsf{T}} w^{\star} \ge w^{\mathsf{T}} w^{\star} + m(w^{\star})$.
(c) The squared Euclidean norm of the iterate satisfies $\|w_u\|^2 \le \|w\|^2 + 1$.
(d) Conclude that perceptron encounters at most $1/m^2(w^{\star})$ misclassification errors.

Solution:

(a) Using expression (60.30) and the fact that $\|w^{\star}\| = 1$, it is clear that $m(w^{\star}) = \min_{0\le n\le N-1} |h_n^{\mathsf{T}} w^{\star}|$.
(b) Note that

$$w_u^{\mathsf{T}} w^{\star} = (w + \gamma h)^{\mathsf{T}} w^{\star} w^{\mathsf{T}} w^{\star} + \gamma h^{\mathsf{T}} w^{\star}$$

Using the definition of the margin we have, for any (h, γ) pair:

$$\gamma h^{\mathsf{T}} w^{\star} = |h^{\mathsf{T}} w^{\star}| \ge m(w^{\star})$$

It follows that

$$w_u^{\mathsf{T}} w^{\star} \ge w^{\mathsf{T}} w^{\star} + m(w^{\star})$$

(c) When a misclassification occurs we have $\gamma h^{\mathsf{T}} w \leq 0$. It follows that

$$\begin{aligned} \|w_u\|^2 &= \|w+\gamma h\|^2 \\ &= \|w\|^2 + \gamma^2\|h\|^2 + 2\gamma h^{\mathsf{T}} w \\ &\leq \|w\|^2 + \gamma^2\|h\|^2 \\ &\leq \|w\|^2 + 1 \end{aligned}$$

(d) Using the results of parts (b) and (c), and after K updates we have

$$w_u^{\mathsf{T}} w^\star \geq Km(w^\star) \quad \text{and} \quad \|w_u\|^2 \leq K$$

and, hence,

$$\begin{aligned} Km(w^\star) &\leq w_u^{\mathsf{T}} w^\star \\ &\leq \|w_u\|\,\|w^\star\| \\ &\leq \|w_u\| \\ &\leq \sqrt{K} \end{aligned}$$

from which we conclude that $K \leq 1/m^2(w^\star)$.

□

5) (**Chapter 61**) Consider a collection of $N = M+1$ feature vectors $\{h_n\}$ in $\mathbb{R}^{M+1}$ with individual entries $\{h_{n,m}\}$ defined by

$$h_{n,m} = \begin{cases} \sqrt{\dfrac{RM}{M+1}}, & \text{when } m = n \\ -\sqrt{\dfrac{R}{M(M+1)}}, & \text{when } m \neq n \end{cases}$$

for some $R > 0$. For instance, the coordinates of the 3 feature vectors in $\mathbb{R}^3$ for $R = 1$ and $M = 2$ are

$$\begin{aligned} h_1 &= \text{col}\left\{\sqrt{\frac{2}{3}},\ -\frac{1}{\sqrt{6}},\ -\frac{1}{\sqrt{6}}\right\} \\ h_2 &= \text{col}\left\{-\frac{1}{\sqrt{6}},\ \sqrt{\frac{2}{3}},\ -\frac{1}{\sqrt{6}}\right\} \\ h_3 &= \text{col}\left\{-\frac{1}{\sqrt{6}},\ -\frac{1}{\sqrt{6}},\ \sqrt{\frac{2}{3}}\right\} \end{aligned}$$

(a) Verify that the feature vectors are centered, i.e.,

$$\bar{h} = \frac{1}{N}\sum_{n=1}^{N} h_n = 0$$

(b) Verify that $\|h_n\|^2 = R^2$ and $h_n^{\mathsf{T}} h_k = -R/M(M+1)$ for any $n \neq k$.

(c) Verify that the given feature vectors lie on the hyperplane that passes through the origin and whose normal direction is the vector $\mathbb{1}_{M+1}$. In other words, the feature vectors lie on a sphere in $\mathbb{R}^M$, which in turn lies on a hyperplane in the higher-dimensional space $\mathbb{R}^{M+1}$.

(d) A random selection of labels $\gamma(n) \in \{\pm 1\}$ is associated with the feature vectors. Refer to the solution of the hard-margin SVM by duality arguments in Section 61.2 in the text. Since the feature vectors are centered, the Lagrangian function (61.34) becomes (by ignoring the offset

parameter θ):

$$\mathcal{L}(w,\lambda(n)) = \frac{1}{2}\|w\|^2 - \sum_{n=0}^{N-1}\lambda(n)\gamma(n)h_n^{\mathsf{T}}w$$

in which case formulation (61.42a)–(61.42b) is replaced by

$$\lambda^\star \;=\; \underset{\lambda\in\mathbb{R}^N}{\text{argmin}}\;\left\{\frac{1}{2}\lambda^{\mathsf{T}}A\lambda \;-\; \mathbb{1}^{\mathsf{T}}\lambda\right\},\quad \text{subject to } \lambda\succeq 0$$

Show that the matrix A can be written in the following form as a rank-one modification of the identity matrix

$$A \;=\; \frac{R}{M(M+1)}\left\{\alpha I_{M+1} - \gamma_N\gamma_N^{\mathsf{T}}\right\}$$

where we introduced the scalar $\alpha = M(M+1)+1$ and $\gamma_N = \text{col}\{\gamma(n)\}$.

(e) Ignore for now the constraint $\lambda\succeq 0$ and show that the $\lambda^\star$ that solves $A\lambda = \mathbb{1}_{M+1}$ is given by

$$\lambda^\star = \frac{\alpha-1}{\alpha R}\left\{\mathbb{1}_{M+1} - \beta\gamma_N\right\}$$

where we introduced

$$\beta = \frac{1}{(M+1)^2+1}\Big(\sum_{n=1}^{M+1}\gamma(n)\Big)$$

Verify that all entries of $\lambda^\star$ are strictly positive.

Solution: This example appears in section 3.3 in the work by Burges, C. (1998), "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167. However, the calculations may not be consistent. We adjust the presentation here for notation and context and provide a different solution method.

(a) Using $N = M+1$, we have

$$\bar{h} \;=\; \frac{1}{M+1}\sum_{n=1}^{M+1}h_n$$

Now, consider an arbitrary mth entry in $\bar{h}$. Then,

$$\begin{aligned}\bar{h}_m \;&=\; \frac{1}{M+1}\sum_{n=1}^{M+1}h_{n,m}\\ &=\; \sqrt{\frac{RM}{M+1}} - M\sqrt{\frac{R}{M(M+1)}}\\ &=\; 0\end{aligned}$$

so that $\bar{h}=0$ and the feature vectors are centered.

(b) Note that for each vector h_n we have

$$\begin{aligned}\|h_n\|^2 &= \frac{RM}{M+1} + M\frac{R}{M(M+1)} \\ &= \frac{RM}{M+1} + \frac{R}{M+1} \\ &= R\end{aligned}$$

Likewise,

$$\begin{aligned}h_n^{\mathsf{T}} h_k &= (M-1) \times \frac{R}{M(M+1)} - \sqrt{\frac{RM}{M+1}}\sqrt{\frac{R}{M(M+1)}} \\ &= \frac{(M-1)R}{M(M+1)} - \frac{R}{M+1} \\ &= (M-1)\frac{R}{M(M+1)} - \frac{MR}{M(M+1)} \\ &= -\frac{R}{M(M+1)}\end{aligned}$$

(c) For any feature vector h_n we have

$$\begin{aligned}\mathbb{1}^{\mathsf{T}} h_n &= \sqrt{\frac{RM}{M+1}} - M\sqrt{\frac{R}{M(M+1)}} \\ &= \sqrt{\frac{RM}{M+1}} - \sqrt{\frac{MR}{(M+1)}} \\ &= 0\end{aligned}$$

Therefore, the entries of the given feature vectors satisfy

$$\sum_{m=1}^{M+1} h_{n,m} = 0$$

so that they lie on a hyperplane whose orthogonal direction is $\mathbb{1}_{M+1}$ and passes through the origin.

(d) The entries of the $(M+1) \times (M+1)$ matrix A are given by

$$[A]_{n,m} = \begin{cases} R, & \text{when } n = m \\ -\dfrac{\gamma(n)\gamma(m)R}{M(M+1)}, & \text{otherwise} \end{cases}$$

That is, the diagonal entries are R while the off-diagonal entries are given by the expression above. We can rewrite A in an alternative form. Introduce the label vector

$$\gamma_N = \text{col}\Big\{\gamma(1), \gamma(2), \ldots, \gamma(N)\Big\}$$

Then,

$$\gamma_N \gamma_N^{\mathsf{T}} = \begin{bmatrix} 1 & \gamma(1)\gamma(2) & \ldots & \gamma(1)\gamma(N) \\ \gamma(2)\gamma(1) & 1 & \ldots & \gamma(2)\gamma(N) \\ \vdots & \cdot & 1 & \vdots \\ \gamma(N)\gamma(1) & \gamma(N)\gamma(2) & \ldots & 1 \end{bmatrix}$$

so that

$$\begin{aligned} A &= \Big(R+\frac{R}{M(M+1)}\Big)I_{M+1}-\frac{R}{M(M+1)}\gamma_N\gamma_N^{\mathsf{T}} \\ &= \frac{R}{M(M+1)}\Big\{\alpha I_{M+1}-\gamma_N\gamma_N^{\mathsf{T}}\Big\} \end{aligned}$$

where we introduced the scalar

$$\alpha = M(M+1)+1$$

We thus note that A has the form of a rank-one modification of the identity matrix. Applying the matrix inversion lemma (29.89) gives

$$A^{-1} = \frac{M(M+1)}{R}\Big\{\frac{1}{\alpha}I_{M+1}-\frac{1}{\alpha}\gamma_N\Big(1+\frac{1}{\alpha}\|\gamma_N\|^2\Big)^{-1}\gamma_N^{\mathsf{T}}\frac{1}{\alpha}\Big\}$$

(e) Next we use $\lambda = A^{-1}\mathbb{1}_{M+1}$ to find (using $\|\gamma_N\|^2 = N = M+1$):

$$\begin{aligned} \lambda^{\star} &= \frac{\alpha-1}{\alpha R}\Big\{\mathbb{1}_{M+1}-\frac{1}{\alpha+M+1}\Big(\sum_{n=1}^{M+1}\gamma(n)\Big)\times\gamma_N\Big\} \\ &= \frac{\alpha-1}{\alpha R}\Big\{\mathbb{1}_{M+1}-\frac{1}{(M+1)^2+1}\Big(\sum_{n=1}^{M+1}\gamma(n)\Big)\times\gamma_N\Big\} \\ &= \frac{\alpha-1}{\alpha R}\Big\{\mathbb{1}_{M+1}-\beta\gamma_N\Big\} \end{aligned}$$

where we introduced the scalar

$$\beta = \frac{1}{(M+1)^2+1}\Big(\sum_{n=1}^{M+1}\gamma(n)\Big)$$

Note that

$$|\beta| \le \frac{\sum_{n=1}^{M+1}|\gamma(n)|}{1+(M+1)^2} = \frac{M+1}{1+(M+1)^2} \le \frac{1}{M+1} < 1$$

It follows that each entry of the vector λ is strictly positive and therefore the constraint $\lambda \succeq 0$ is satisfied. On the other hand, we have

$$\begin{aligned} \gamma_N^{\mathsf{T}}\lambda &= \frac{\alpha-1}{\alpha R}\Big\{\gamma_N^{\mathsf{T}}\mathbb{1}_{M+1}-\beta\|\gamma_N\|^2\Big\} \\ &= \frac{\alpha-1}{\alpha R}\Big\{\sum_{n=1}^{M+1}\gamma(n)-\beta(M+1)\Big\} \\ &= \frac{\alpha-1}{\alpha R}\Big\{\sum_{n=1}^{M+1}\gamma(n)-\frac{M+1}{(M+1)^2+1}\Big(\sum_{n=1}^{M+1}\gamma(n)\Big)\Big\} \end{aligned}$$

□

6) (**Chapter 63**) The ℓ_2-regularized logistic regression problem involves minimizing the following empirical risk over the training dataset $\{\gamma(n), h_n\}$:

$$\min_{w\in\mathbb{R}^M}\Big\{\frac{\rho}{2}\|w\|^2 \;+\; \frac{1}{N}\sum_{n=0}^{N-1}\ln\Big(1+e^{-\gamma(n)\widehat{\gamma}(n)}\Big)\Big\}$$

where $\widehat{\gamma} = h^{\mathsf{T}} w$. In the above formulation we are assuming, for simplicity, that the feature data has been centered so that the offset parameter θ can be set to 0 and it is sufficient to seek a weight vector w to separate the classes $\gamma \in \{\pm 1\}$. In this problem we wish to examine a kernal-based formulation of logistic regression. Introduce a kernel function $K(h_a, h_b)$ and the corresponding $N \times N$ Gramian matrix

$$A_{m,m'} = K(h_m, h'_m), \quad m, m' = 0, 1, \ldots, N-1$$

(a) Argue that a kernelized version of logistic regression requires solving

$$\min_{\alpha \in \mathbb{R}^N} \mathcal{P}(\alpha) \;\triangleq\; \left\{ \frac{\rho}{2} \alpha^{\mathsf{T}} A \alpha \;+\; \frac{1}{N} \sum_{n=0}^{N-1} \ln\Big(1 + e^{-\gamma(n)[A\alpha]_n}\Big) \right\}$$

where the notation $[x]_n$ extracts the nth entry of vector x.

(b) Let $\sigma(z) = \ln(1 + e^{-z})$. Verify that the gradient vector of $\mathcal{P}(\alpha)$ relative to α is given by

$$\nabla_{\alpha^{\mathsf{T}}} \mathcal{P}(\alpha) = \rho A \alpha + \frac{1}{N} A D(\alpha) \gamma_{\text{vec}}$$

where $\gamma_{\text{vec}} = \text{col}\{\gamma(0), \ldots, \gamma(N-1)\}$ and $D(\alpha) = \text{diag}\Big\{\sigma'\Big(\gamma(n)[A\alpha]_n\Big)\Big\}$.

(c) Write down a stochastic gradient recursion for determining the minimizer $\alpha^\star$ to the problem in part (a).

Solution:

(a) We let $\mathcal{H}$ denote the reproducing kernel Hilbert space (RKHS) associated with the kernel function $K(h, h')$. This is a space of functions of the feature variable, h. For example, the prediction $\widehat{\gamma}(h)$, which is a function of h, lives in this space. In a manner similar to (63.80) in the body of the chapter, we then consider the problem

$$\widehat{\gamma}^\star(h) \;\triangleq\; \underset{\widehat{\gamma}(h) \in \mathcal{H}}{\text{argmin}} \left\{ \frac{\rho}{2} \|\widehat{\gamma}(h)\|_{\mathcal{H}}^2 \;+\; \frac{1}{N} \sum_{m=0}^{N-1} \ln\Big(1 + e^{-\gamma(n)\widehat{\gamma}(h_n)}\Big) \right\}$$

where $\rho > 0$ is the regularization parameter. We know from the Representer theorem that the optimal solution) has the form

$$\widehat{\gamma}^\star(h) \;=\; \sum_{m=0}^{N-1} \alpha^\star(m) K(h, h_m)$$

for some real coefficients $\{\alpha^\star(m)\}$. This motivates us to proceed as follows. Let $\alpha \in \mathbb{R}^{N \times 1}$ denote a column vector of size N. For any feature vector h, we introduce the following column vector involving kernel evaluations of h with the training data:

$$u_h \;\triangleq\; \text{col}\Big\{K(h, h_0),\, K(h, h_1),\, \ldots,\, K(h, h_{N-1})\Big\} \in \mathbb{R}^{N \times 1}$$

We also introduce the $N \times N$ Gramian matrix

$$[A]_{m,m'} \;\triangleq\; K(h_m, h_{m'}), \;\; m, m' = 0, 1, \ldots, N-1$$

and the matrix of transformed feature vectors

$$\Phi \;\triangleq\; \big[\; h_0^\phi \;\; h_1^\phi \;\; \ldots \;\; h_{N-1}^\phi \;\big] \in \mathbb{R}^{M_\phi \times N}$$

so that $A = \Phi^{\mathsf{T}} \Phi$. The nth column of A corresponds to the kernel evaluations of h_n with all

other training vectors. We denote this column by u_n so that

$$u_n \triangleq u_{h_n} = n\text{th column of } A = \begin{bmatrix} K(h_n,h_0) \\ K(h_n,h_1) \\ \vdots \\ K(h_n,h_{N-1}) \end{bmatrix} = \begin{bmatrix} (h_0^\phi)^\mathsf{T} h_n^\phi \\ (h_1^\phi)^\mathsf{T} h_n^\phi \\ \vdots \\ (h_{N-1}^\phi)^\mathsf{T} h_n^\phi \end{bmatrix} = \Phi^\mathsf{T} h_n^\phi$$

The Representer theorem shows that we can parameterize the sought-after function $\widehat{\gamma}(h)$ in the following *linear* form in the expanded domain:

$$\widehat{\gamma}(h) = u_h^\mathsf{T}\alpha = (h^\phi)^\mathsf{T}\Phi\alpha = [A\alpha]_n$$

for some vector $\alpha \in \mathbb{R}^N$. Likewise, we can replace the regularization factor $\|\widehat{\gamma}(h)\|_{\mathcal{H}}^2$) by the quadratic form

$$\|\widehat{\gamma}(h)\|_{\mathcal{H}}^2 = \alpha^\mathsf{T} A\alpha = \alpha^\mathsf{T}\Phi^\mathsf{T}\Phi\alpha = \alpha^\mathsf{T} A\alpha$$

Substituting these definitions into the kernelized empirical risk we get

$$\alpha^\star \triangleq \underset{\alpha\in\mathbb{R}^N}{\text{argmin}}\ \mathcal{P}(\alpha) \triangleq \left\{ \frac{\rho}{2}\alpha^\mathsf{T} A\alpha + \frac{1}{N}\sum_{m=0}^{N-1} \ln\Big(1+e^{-\gamma(n)[A\alpha]_n}\Big) \right\}$$

(b) Let $\sigma(z) = \ln(1+e^{-z})$. Then, $\sigma'(z) = -1/(1+e^z)$. Computing the derivative of $\mathcal{P}(\alpha)$ relative to the ℓth entry of α gives

$$\frac{\partial\mathcal{P}(\alpha)}{\partial\alpha_\ell} = \rho[A\alpha]_\ell + \frac{1}{N}\sum_{n=0}^{N-1}\gamma(n)A_{n\ell}\times\sigma'\Big(\gamma(n)[A\alpha]_n\Big)$$

Grouping terms we obtain

$$\nabla_{\alpha^\mathsf{T}}\mathcal{P}(\alpha) = \rho A\alpha + \frac{1}{N}AD(\alpha)\gamma_{\text{vec}}$$

(a) It follows that we can use the update

$$\alpha_m = \alpha_{m-1} - \mu\nabla_{\alpha^\mathsf{T}}\mathcal{P}(\alpha_{m-1})$$

□

7) (**Chapter 64**) Let $\mathcal{C}$ denote the set of all possible affine classifiers ($w\in\mathbb{R}^M, \theta\in\mathbb{R}$) that can be generated by the perceptron algorithm when applied to a linearly separable training dataset. Let $\mathcal{M}$ denote the maximum number of misclassifications that the algorithm can encounter. Show that the VC dimension of the set $\mathcal{C}$ satisfies

$$\text{VC}(\mathcal{C}) \le \min\Big\{\mathcal{M}, M+1\Big\}$$

Solution: The VC dimension of affine classifiers is $M+1$. Since $\mathcal{C}$ is a subset of the collection of affine classifiers, then we have that $\text{VC}(\mathcal{C}) \le M+1$. Let $\{h_1, h_2, \ldots, h_{M+1}\}$ denote a set of feature vectors that can be shattered by $\mathcal{C}$. This means that for any randomly selected labels $\{\gamma(1),\gamma(2),\ldots,\gamma(M+1)\}$, there will exist a classifier $w^\star\in\mathcal{C}$ such that

$$\gamma(n)h_n^\mathsf{T} w^\star > 0, \quad n=1,2,\ldots,M+1$$

Here we are assuming the weight vector and the feature vectors are extended according to (60.20) for convenience. We apply the perceptron algorithm to determine a separating hyperplane from within the set $\mathcal{C}$ for this collection. We can construct a situation where the algorithm can make $M+1$

mistakes during this run so that

$$\#\text{ of mistakes} \geq \text{VC}(\mathcal{C})$$

Indeed, at any iteration $1 \leq n \leq M+1$, the perceptron algorithm generates a predicted label $\widehat{\gamma}(n)$. This label may agree with $\gamma(n)$ or it may be wrong. If it agrees with $\gamma(n)$, then we could select a different model from within $\mathcal{C}$ that generates instead $-\gamma(n)$ as the label for h_n. This is possible since the $\{h_1, h_2, \ldots, h_{M+1}\}$ are shattered by $\mathcal{C}$. In this way, we end up having $M+1$ mistakes. We know that the number of mistakes can never exceed $\mathcal{M}$. Then, we get $\text{VC}(\mathcal{C}) \leq \mathcal{M}$ and the result follows.

□

8) (**Chapter 65**) Consider the 3-node neural network shown in Fig. 1. The output node is simply an adder providing

$$\widehat{\gamma} = w_3 y_1 + w_4 y_2$$

where $\{y_1, y_2\}$ are the outputs of the internal nodes with ReLu activation functions, i.e.,

$$y_1 = \max\{0, hw_1 - \theta_1\}, \qquad y_1 = \max\{0, hw_2 - \theta_2\}$$

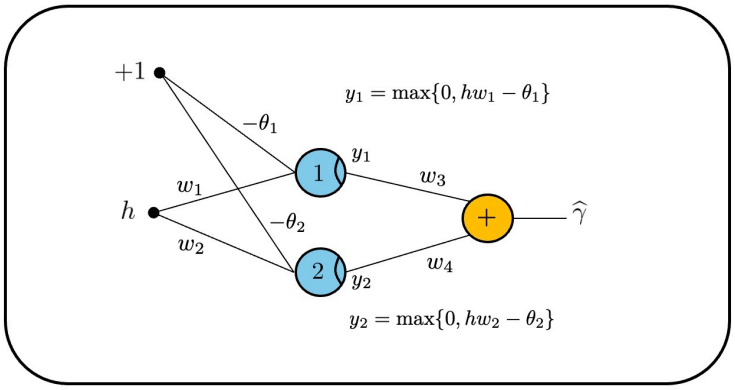

Fig. 1. A neural network with three nodes and ReLu activation functions in the first two nodes. The output node is an adder.

The input is a scalar feature, $h \in \mathbb{R}$, and the output is another scalar $\widehat{\gamma} \in \mathbb{R}$. Therefore, the network implements an input–output mapping from h to $\widehat{\gamma}$, denoted by $\widehat{\gamma} = \mathcal{N}(h; P)$. The mapping is characterized by 6 scalar parameters collected into P:

$$P \triangleq \{w_1, w_2, w_3, w_4, \theta_1, \theta_2\}$$

(a) Verify that the following 2 choices of parameters lead to the same output value, i.e., $\mathcal{N}(h; W_1) = \mathcal{N}(h; W_2)$, where

$$\begin{aligned} W_1 &= \{1, 1, 1, -1, -1, 0\} \\ W_2 &= \{1, 1, -1, 1, 0, -1\} \end{aligned}$$

(b) Consider the following convex combination of the above 2 sets of parameters,

$$W_3 = \frac{1}{2}W_1 + \frac{1}{2}W_2 = \{1, 1, 0, 0, -1/2, -1/2\}$$

Verify that now $\mathcal{N}(h; W_3) = 0$ independent of the value of h.

(c) Consider the least-squares empirical risk

$$R_{\text{emp}}(W) = \frac{1}{N}\sum_{n=0}^{N-1}(\gamma(n) - \widehat{\gamma}(n))^2$$

Consider $N = 2$ with data $(h(0), \gamma(0)) = (-1, 1)$ and $(h(1), \gamma(1)) = (1, -1)$. Evaluate the empirical risk for the choices W_1, W_2, and W_3. Does the empirical risk depend on the parameters W in a convex manner?

Solution:

(a) For the first set of parameters we get

$$\begin{aligned} \widehat{\gamma} &= w_3 y_1 + w_4 y_2 \\ &= y_1 - y_2 \\ &= \max\{0, h+1\} - \max\{0, h\} \end{aligned}$$

For the second set of parameters we get

$$\begin{aligned} \widehat{\gamma} &= w_3 y_1 + w_4 y_2 \\ &= -y_1 + y_2 \\ &= -\max\{0, h\} + \max\{0, h+1\} \end{aligned}$$

so that $\mathcal{N}(h; W_1) = \mathcal{N}(h; W_2)$.

(b) We now have

$$\widehat{\gamma} = w_3 y_1 + w_4 y_2 = 0$$

(c) For W_1 we have $\gamma(0) = 1$ and $h(0) = -1$:

$$\begin{aligned} \widehat{\gamma}(0) &= \max\{0, h(0)+1\} - \max\{0, h(0)\} \\ &= \max\{0, 0\} - \max\{0, -1\} \\ &= 0 \end{aligned}$$

and for $\gamma(1) = -1$ and $h(0) = 1$ we have

$$\begin{aligned} \widehat{\gamma}(1) &= \max\{0, h(1)+1\} - \max\{0, h(1)\} \\ &= \max\{0, 2\} - \max\{0, 1\} \\ &= 1 \end{aligned}$$

It follows that

$$R_{\text{emp}}(W_1) = \frac{1}{2}\Big(1^2 + 2^2\Big) = 5/2$$

For W_2, the input-output map is the same and therefore we also get

$$R_{\text{emp}}(W_2) = 5/2$$

On the other hand, for W_3 we have $\widehat{\gamma} = h$ independent of h and, therefore,

$$R_{\text{emp}}(W_3) = \frac{1}{2}\Big(2^2\Big) = 2$$

It is clear that

$$R_{\text{emp}}(W_3) \neq \frac{1}{2} R_{\text{emp}}(W_1) + \frac{1}{2} R_{\text{emp}}(W_2)$$

which illustrates that the empirical risk does not depend on the parameters in a convex manner.

□