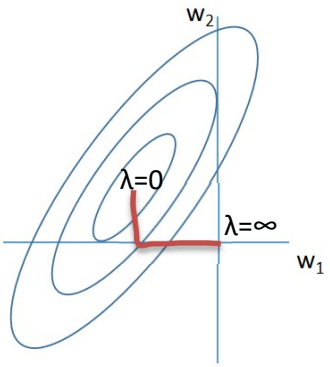
EE 769 Introduction to Machine Learning (IIT Bombay)

Mid-Semester Examination Solution Guide (2023.02.23 :: 0830—1030) [20 points]

1. **L1 regularization:** Contours of an unregularized convex training loss function with respect to the parameters w_1 and w_2 are shown in the figure on the right.



   a. Trace the approximate locus of the optimal weights as the L1 regularization penalty on the weight vector $[w_1\ w_2]^T$ is increased from 0 to ∞, clearly indicating the least and the most regularized solutions. Assume that the innermost contour is tangential to w_1 axis, the middle contour is tangential to the w_2 axis and intersects w_1 axis at 45^O and 80^O, and the outermost contour intersects w_1 axis at 60^O and w_2 axis at 70^O and w_2 axis at 70^O and 20^O. All angles are measured counterclockwise with respect to the positive w_1 axis. [2]

For λ = 0, the center of the ellipse will be solution [0.5].
For λ=∞, the origin will be the solution [0.5].
At an intermediate λ, the locus will touch the w_1 axis at the middle contour, where the angle is 45^0 [0.5], after which for larger λ w_2 will remain 0 [0.5].

   b. Write a general loss function for mean square error with L1 regularization penalty on the weight vector assuming two weights, one bias, and N training samples. Your final answer should be in terms of target values, weights and biases. [1] ***1/2N $\sum_i (w^T x_i + b - t_i)^2 + \lambda \sum_j |w_j|$*** **[0.5 + 0.5]**

   c. Given two regularization penalties $\lambda_1$and λ_2, how will you determine which is better in a general scenario? [1] ***The one that gives better validation performance on a set-aside validation sample.***

2. **Bayesian classification:** Two class conditional densities are given by the following expressions:

$$p(x|c_0) = \begin{cases} k_0\sqrt{4-x^2}, \text{if } -2 < x < 2 \\ 0, \text{otherwise} \end{cases}, p(x|c_1) = \begin{cases} k_1\sqrt{9-(x-4)^2}, \text{if } 1 < x < 7 \\ 0, \text{otherwise} \end{cases}.$$

   a. What is the decision boundary if $(c_0) = \frac{4}{13}$? (Hint: Determine k_0/k_1 first, which is easy.) [3]

At the decision boundary p(x|c_0)p(c_0) / p(x|c_1) / p(c_1) =1. [0.5]
Because probability densities should integrate to 1, k_0/k_1 = 9/4. [1]
Because priors add up to 1, p(c_1) = 9/13. [0.5]
Now, all that remains is $\sqrt{4-x^2} = \sqrt{9-(x-4)^2}$ => $x^2 = (x-4)^2 - 5 = x^2 - 8x + 11$ => $= 11/8$. [1]

   b. For two overlapping uniform class conditional densities of a single variable, p(**x**|c=0) is non-zero from *x* = *a* to *x* = *b*, and p(**x**|c=1) is non-zero from *x* = *c* to *x* = *d*. Assume *a* < *c* < *b* < *d*. For what value of the p(c=0) is it difficult to find the decision boundary threshold on *x* axis for a Bayesian decision criterion? [2]

For uniform distributions p(x|c=0) = 1/(b − a) between a and b, and p(x|c=1) = 1/(d − c) between c and d [0.5]
At the decision boundary p(x|c=0)p(c=0) = p(x|c=1) p(c=1), and p(c=1) = 1 − p(c=0). [0.5],
Due to uniform distribution, when the two sides are equal, we will get an ambiguous region between c and b [0.5],
Therefore, for indecision p(c=0) / (b − a) = (1 − p(c=0)) / (d − c) => p(c=0) = (b − a) / (d − c + b − a) [0.5]

3. **Feature engineering:** A linear binary classifier needs to be trained on a subset of four variables using feature selection methods. Assume that the criteria used is training accuracy (proportion of correctly classified samples) for subset evaluation. Based on the figures showing the training data for two pairs of variables (assume no good relation between the pairs not shown), answer the following:

   a. Write a general pseudo-algorithm for forward selection. [1.5]

Let S_0 be the set of selected variables, which is an empty set initially
Let V be the set of remaining variables, which is all variables initially [0.5]
For i = 1 to d
    $fitness_{max}$ = −∞
    For each variable v in V [0.5]
        fitness = fitness of S_{i-1} + {v}
        If fitness > $fitness_{max}$
            v_{max} = v

$fitness_{max}$ = fitness

$S_i = S_{i-1} + \{v_{max}\}$ [0.5]

$V = V - \{v_{max}\}$

b. Write a general pseudo-algorithm for backward elimination. [1.5] WRITE FINAL ANSWER HERE:

Let S_0 be the set of selected variables, which is all variables initially

Let V be the set of discarded variables, which is an empty set initially [0.5]

For i = 1 to d

$fitness_{max} = -\infty$

For each variable v in V [0.5]

fitness = fitness of $S_{i-1} - \{v\}$

If fitness > $fitness_{max}$

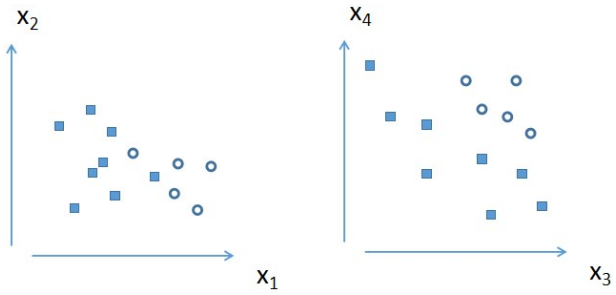
$v_{max} = v$

$fitness_{max}$ = fitness

$S_i = S_{i-1} - \{v_{max}\}$ [0.5]

$V = V + \{v_{max}\}$

c. What will be sequence of inclusion of variables when forward selection is used on the data shown? Justify. [1]



The first to be included will be x_1, as it has a threshold that has the lowest misclassification (only 1). [0.5]

The last to be included will be x_2, as its best threshold has the highest misclassification. [0.5]

d. What will be the sequence of exclusion of variables when backward elimination is used on the data shown? Justify. [1]

The first to be excluded will be x_1 or x_2, as a linear classifier on x_3 and x_4 can perfectly separate training points. [1]

e. Assume a data matrix $\boldsymbol{X}$ for which $x_{i,j}$ is the j^{th} dimension of the i^{th} sample, write the steps to obtain a data matrix $\widehat{\boldsymbol{X}}$ with mean and variance normalized. [1]

Compute mean and variance dimension-wise, i.e. $\mu_j = 1/N \sum_i x_{ij}$ and $\sigma_j^2 = 1/N \sum_i (x_{ij} - \mu_j)^2$ [0.5]

From each column, compute normalized variable , i.e. $\hat{x}_{ij} = (x_{ij} - \mu_j) / \sigma_j$. [0.5]

4. **Linear SVM:**

a. Show that the following two optimization problems for hard SVM will lead to the same solution [2.5]:

i. $\max_{\mathbf{w},b} [\min_i ||\mathbf{w}^T\mathbf{x}_i+b||]$, subject to: $||\mathbf{w}|| = 1$ and $\forall i$, $t_i (\mathbf{w}^T\mathbf{x}_i+b) \geq 0$;

ii. $\min_{\mathbf{v},c} ||\mathbf{v}||^2$, subject to: $\forall i$, $t_i (\mathbf{v}^T\mathbf{x}_i+c) \geq 1$

Let $\mathbf{v} = \alpha\mathbf{u}$, where $\alpha = ||\mathbf{v}|| \geq 0$. [0.5]

Thus, the second optimization problem becomes $\max_{\alpha,u,c} 1/\alpha$, s.t. $||\mathbf{u}|| = 1$ and $\forall i$, $t_i (\mathbf{u}^T\mathbf{x}_i+c) \geq 1/\alpha$ [0.5]

When $||\mathbf{u}|| = 1$, then $||\mathbf{u}^T\mathbf{x}_i+c||$ is distance of $\mathbf{x}_i$ from the hyperplane defined by u and c [0.5]

Thus, $[\min_i ||\mathbf{u}^T\mathbf{x}_i+c||] = 1/\alpha$. [0.5]

Thus, the objective becomes $\max_{u,c}$ $[\min_i ||\mathbf{u}^T\mathbf{x}_i+c||]$, s.t. $||\mathbf{u}|| = 1$ and $\forall i$, $t_i (\mathbf{u}^T\mathbf{x}_i+c) \geq 0$ [0.5]

b. A soft margin SVM trained using the optimization problem of the form $\min_{\mathbf{v},c} ||\mathbf{v}||^2$, subject to: $\forall i$, $t_i (\mathbf{v}^T\mathbf{x}_i+c) \geq 1$. Fill the last three columns of the table below (S.V. = support vec.; two col.s are yes/no) . [2.5]

Point $\mathbf{x}_i$	Target t_i	Model $\mathbf{v}^T\mathbf{x}_i+c$	Misclassified?	S.V.?	Reason
$\mathbf{x}_1$	+1	−1	Yes	Yes	Misclassified points are SV; their contribution to the solution is non-zero
$\mathbf{x}_2$	+1	+1	No	Yes	On-margin points are SV; their contribution to the solution is non-zero
$\mathbf{x}_3$	−1	−2	No	No	Points away from the margin are not SV
$\mathbf{x}_4$	−1	+2	Yes	Yes	Misclassified points are SV; their contribution to the solution is non-zero
$\mathbf{x}_5$	−1	0	Undecided	Yes	Points inside the margin are SV; their contribution to the solution is > NZ

Give points only if both the columns -- misclassified and S.V. -- are correct, and some coherent reason is given.