1 Standard Derivations

1. Consider the PCA set-up where dimensionality has to be reduced from n to unity. More specifically, let the encoding of $x \in \mathbb{R}^n$ be $w^\top x$, where $w \in \mathbb{R}^n$ is the parameter to be learnt from the training data: $\mathcal{D} = \{x_1, \ldots, x_m\}$. The stochastic optimization problem formalizing the goal of minimizing reconstruction-error in this case is:

$$\min_{w \in \mathbb{R}^n, v \in \mathbb{R}^n} \mathbb{E}_{X \sim p^*} \underline{\left[\|X - v w^T X\|^2 \right]}, \quad \text{——— ①}$$

where p^* is the underlying likelihood function, whose samples are in $\mathcal{D}$ (unsupervised learning assumption) and v is the parameter of the decoding model [[*Fill in the above blank with an appropriate mathematical expression involving* v, w, X. 1 **Mark**]].

Using simple linear algebra (least-squares solution style) arguments show that, without loss of generality, one can restrict $v = w, \|w\| = 1$ in the above problem. Note down these arguments in the box below:

$w^T X$ is some number. Let's ask which number is "best" $\rightarrow$ ~~decoding~~ corresponding to v. $\arg\min_{\alpha} \|X - v\alpha\|^2 = \alpha = \frac{v^T X}{\|v\|^2}$.	① $\geq \min_{v} E\left[\left\|X - \frac{v v^T X}{\|v\|^2}\right\|^2\right]$ $= \min_{\|v\|=1} E[\|X - v v^T X\|^2]$ But trivially ① $\leq \nearrow$ $\therefore$ ① $= \circlearrowleft$

$\rightarrow$ See Lemma 23.1 in "Understanding ML" textbook for more details, which is cited in summary notes.

[[*writings outside the box, and illegible writings, will be strictly ignored by the evaluator* 1 **Mark**]]

Now perform algebraic simplifications to show that the above stochastic optimization problem is equivalent to:

$$\max_{\|w\|=1} w^\top \underline{E_{X \sim p^*}\left[X X^T \right]} w.$$

[[*Fill in the above blank with an appropriate mathematical expression involving* X, p^*. 0.5 **Marks**]]. The SAA (ERM) version of the above is:

$$\max_{\|w\|=1} w^\top \underline{\frac{1}{m} \sum_{i=1}^{m} x_i x_i^T} w.$$

[[*Fill in the above blank with an appropriate mathematical expression involving* $x_1, \ldots, x_m$. 0.5 **Marks**]]. Now, let the $n \times n$ matrix in the above blank be denoted by M. Let $u_1, \ldots, u_n$ be the eigen vectors corresponding to the eigenvalues of M, which are $\lambda_1, \ldots, \lambda_n$ written in decreasing order. Recall from spectral theorem that $u_1, \ldots, u_n$ can be chosen to be unit vectors orthogonal to each other. In particular, they form an orthogonal basis for R^n and hence any $w \in \mathbb{R}^n$ can be re-parametrized as $U\alpha$ (change of variables), where U is the orthogonal matrix whose columns are $u_1, \ldots, u_n$. With this change of variables, the above optimization problem simplifies as:

Details are in Theorem 23.2 in same book and done in lecture.

$$\max_{\|\alpha\|=1} \underline{\sum_{i=1}^{n} \lambda_i \alpha_i^2}.$$

[[*Fill in the above blank with an appropriate mathematical expression involving* $\lambda_1, \ldots, \lambda_n$ *and entries of* α, *which are* $\alpha_1, \ldots, \alpha_n$. 1 **Mark**]]. By inspection, it is clear that the optimal solution of this problem is $\alpha^* = \begin{bmatrix} \underline{1} \\ \underline{0} \\ \vdots \\ \underline{0} \end{bmatrix}$. [[*Fill in these blanks with*

(2)

appropriate numbers. 0.5 **Marks**]]. Hence, the optimal $w^* =$ u_1. [[*Fill in the blank with an appropriate mathematical expression involving one or few of* $u_1, \ldots, u_n$. 0.5 **Marks**]].

2. Consider the ERM problem with l_2 regularized linear models and a loss function, l:

$$\min_{w \in \mathbb{R}^n} \quad \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{m} l\left(w^\top \phi(x_i), y_i\right)$$

Show that any optimal solution of this problem can be written as a linear combination of the training datapoints in the feature space. Note your proof in the box below:

Any $w \in \mathbb{R}^n$ can be written as lin. comb of $\varphi(x_1) \ldots \varphi(x_m) \rightarrow \sum_{i=1}^{m} \alpha_i \varphi(x_i)$ plus $w_\perp \in \mathbb{R}^n \ni w_\perp \perp \varphi(x_i)\ \forall\ i$	First term $= \frac{1}{2}\|\sum \alpha_i \varphi(x_i)\|^2 + \frac{1}{2}\|w_\perp\|^2$ by pythagorous theorem. Second term does NOT involve $w_\perp$ $\therefore\ w_\perp = 0$. (soln. of $\underset{w_\perp}{\text{argmin}} \|w_\perp\|^2$).

[[*writings outside the box, and illegible writings, will be strictly ignored by the evaluator* 2 **Marks**]]
Using this, the above optimization problem can be re-written as:

α_i is wrong 0 marks.

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j \varphi(x_i)^\top \varphi(x_j) + C\sum_{i=1}^{m} l(y_i, \sum_{j=1}^{m} \alpha_j \varphi(x_j)^\top \varphi(x_i)).$$

[[*Fill in these blanks with appropriate expressions involving entries of* α *and dot-products between training datapoints in the feature space.* 0.5+0.5=1 **Mark**]].

3. In this question you must re-derive the MCLE problem from first principles. Recall that the goal in training with the discriminative models is to find a $q \in \mathcal{Q}$ such that the corresponding posterior $q(y/x)$ is close to the true posterior $p^*(y/x)$ for typical inputs from $p^*(x)$. This goal is formalized in the following problem, using a loss l between likelihood functions:

See Section 2.3 in Summary Notes for details.

$$\arg\min_{q \in \mathcal{Q}} \mathbb{E}_{X \sim p^*(x)} \left[l(p^*(\cdot/X), q(\cdot/X)) \right],$$

$\rightarrow \mathcal{X}\ l(p^*(y/x), q(y/x))$

[[*Fill in the blank with appropriate expression involving* l, p^*, q, X. 0.5 **Marks**]].
When l is KL divergence, the above can be re-written as:

$$\arg\min_{q \in \mathcal{Q}} \mathbb{E}_{X \sim p^*(x)} \left[\mathbb{E}_{Y/X \sim p^*(Y/X)} \left[\log\left(\frac{p^*(Y/X)}{q(Y/X)} \right) \right] \right],$$

[[*Fill in the blank with appropriate expression involving* p^*, q, X, Y. 0.5 **Marks**]].
Using total expectation rule, the above simplifies as the following stochastic optimization problem given the training set [[*Fill the earlier blank with the name of an appropriate standard result in probability theory* 0.5**Marks**]]:

$$\arg\max_{q \in \mathcal{Q}} \mathbb{E}_{(X,Y) \sim p^*(x,y)} \left[\log(q(Y/X)) \right].$$

$\rightarrow q(y/x)$ wrong
$\rightarrow q(y/x)$ wrong.

[[*Fill in the blank with an appropriate expression involving* q, X, Y. 0.5 **Marks**]].
The SAA version of this problem is the MCLE problem given by:

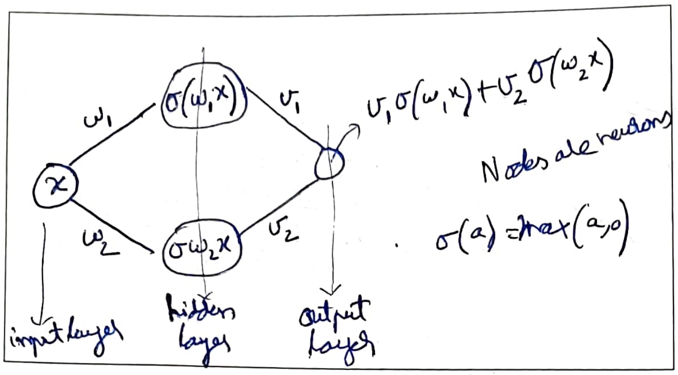
$$\arg\max_{q \in \mathcal{Q}} \frac{1}{m}\sum_{i=1}^{m} \left[\log(q(y_i/x_i)) \right].$$

$\rightarrow q(y/x)$ wrong.

[[*Fill in the blank with an appropriate expression involving the training datapoints* $(x_1, y_1), \ldots, (x_m, y_m)$ *and* q. 0.5 **Marks**]].

2 Simple Exercises

1. Consider a regression problem where the input space as well as the label space is the set of real numbers. Consider a FFNN model appropriate for this problem with width=2, depth=1, and activation function as ReLU. Let loss be the squared-loss. In the box below, draw an illustration of the FFNN. Clearly mark out the artificial neurons, input,hidden,output neural layers, edge weights (parameters), activations at various neurons in response to some input x [[1 Mark]].



In the box below, write down the ERM problem using the notation in your illustration, and assuming the training data is $\{(1,1),(2,2)\}$. Write a simplified expression. [[0.5 Mark]]:

$$\min_{w_1,w_2,v_1,v_2\in\mathbb{R}} \left(\left(v_1\sigma(w_1)+v_2\sigma(w_2)-1\right)^2 + \left(v_1\sigma(2w_1)+v_2\sigma(2w_2)-2\right)^2 \right)$$

In the box below, write down the output of the backprop algorithm run on this network&data, as analytical expressions in terms of the parameter values at the start of backprop [[*Please do not write any details/derivation/calculations. Only write the final expressions denoting the output of backprop.* 1 Mark]].

Let $g(v_1,v_2,w_1,w_2) \equiv \left(v_1\sigma(w_1)+v_2\sigma(w_2)-1\right)^2$.

$$\nabla g\Big|_{\substack{v_1=v_1^0\\ v_2=v_2^0\\ w_1=w_1^0\\ w_2=w_2^0}} = \begin{bmatrix} a\,\sigma(w_1^0) \\ a\,\sigma(w_2^0) \\ v_1^0 a\, 1_{\{w_1^0>0\}} \\ v_2^0 a\, 1_{\{w_2^0>0\}} \end{bmatrix}, \text{ where } a \equiv \underbrace{2\left(v_1^0\sigma(w_1^0)+v_2^0\sigma(w_2^0)-1\right)}_{\text{Computed by forward pass.}}$$

before back prop.

2. Consider a binary classification problem with inputs in $\mathbb{R}^2$. Consider kernels k_1, k_2 over $\mathbb{R}^2$ defined by $k_1(x,y) \equiv x^\top y, k_2(x,y) \equiv 1 + x^\top y$. Let the training data be $D = \left\{ \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 2 \\ 2 \end{bmatrix}, -1 \right) \right\}$. Then the optimal solution of the corresponding ERM

problem with hard-margin SVM and kernel k_1 is given by: DOES NOT EXIST [[1 Mark]]. Solve the corresponding ERM problem with hard-margin SVM and kernel k_2 and note the important steps in your derivation in the box below. Highlight the final optimal solution [[3 Marks]]:

ERM is $\min_{\alpha_1 \in \mathbb{R}, \alpha_2 \in \mathbb{R}} \frac{1}{2} [\alpha_1 \ \alpha_2] \begin{bmatrix} 3 & 5 \\ 5 & 9 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$

s.t. $1(3\alpha_1 + 5\alpha_2) \geq 1$

$-1(5\alpha_1 + 9\alpha_2) \geq 1$

wrt α_1 is $\min_{\alpha_1 \in \mathbb{R}} 3\alpha_1^2 + 10\alpha_1\alpha_2 \quad \Rightarrow \alpha_1^* = \frac{1-5\alpha_2}{3}$.

s.t. $\frac{1-5\alpha_2}{3} \leq \alpha_1 \leq \frac{-9\alpha_2 - 1}{5}$

eliminating α_1 gives $\min_{\alpha_2 \in \mathbb{R}} 3\left(\frac{1-5\alpha_2}{3}\right)^2 + 10\left(\frac{1-5\alpha_2}{3}\right)\alpha_2 + 9\alpha_2^2$

s.t. $\frac{1-5\alpha_2}{3} \leq \frac{-9\alpha_2 - 1}{5}$

$\alpha_2^* = -4 \Rightarrow \alpha_1^* = 7$

Full marks even if you follow Method ① in rough sheets.

Hint: *Optimization with multiple variables is similar to integration with multiple variables; can be simplified by successive elimination. Alternatively, use geometric insights taught in lecture about the optimal solution.*

The predicted score (un-thresholded) with this optimal solution for the input $\begin{bmatrix} 2 \\ 3 \end{bmatrix}$ is -2. [[*Fill in this blank with a number.* 1 Mark]].

3. Consider an n-dimensional k-component Gaussian Mixture Model (GMM). The likelihood functions in this model are given by the expression

$$p(x) = \sum_{i=1}^{k} \frac{\theta_i}{(2\pi)^{n/2} |\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)}.$$

equation (3.9) in Summary Notes

In this expression, $\theta_1 \ldots \theta_k, \mu_1 \ldots \mu_k, \Sigma_1 \ldots \Sigma_k$ denote the parameters of the GMM. [[1+0.5=1.5 Marks]]. While employing GMM for clustering one makes additional assumptions, which provide a relevant meaning to these parameters. In the box below recall these important assumptions [[1.5 Marks]]:

$Y = \{1 \ldots k\} \rightarrow$ cluster Ids

$\exists\; p^*(x,y) = p^*(x/y)\, p^*(y)$

$\ni p^*(x/y) \rightarrow$ Gaussian $\forall y$

Training set is iid samples from

$p^*(x) \equiv \sum_y p^*(x,y)$ (marginal)

$\longrightarrow$ Sec 5.19 in Summary Notes

Under what (theoretical/asymptotic) conditions on the training data and the training algorithm is exact recovery of the likelihood functions that appear in these assumptions possible? List these in the following box [[2 Marks]]:

(1) no. Samples $\rightarrow \infty$

(2) MLE solver finds "global" optimal parameters. $\longrightarrow$ Challenging as non-convex problem.

3 True-or-False Type Questions

NOTE: Fill in the blanks in this section appropriately with either "TRUE" or "FALSE". Each blank carries +0.5 marks when correctly answered and carries -0.1 marks when answered wrongly. So beware of the NEGATIVE marking. It is recommended you attempt a question ONLY IF you are sure about the correctness of your answer.

1. Consider the online learning set-up where there exists a $f^* \in \mathcal{F}$ such that all the examples (x, y) satisfy $f^*(x) = y$. Let $|\mathcal{F}| = 8$ and number of examples is $m = 10$. Then, an upper bound on the number of mistakes the `halving algorithm` makes in the worst-case in $m = 10$ rounds is 3. TRUE.

2. Consider a special case where the number of clusters obtained with a 5-component Gaussian mixture model is 5. Then, the shape of these clusters will be elliptical. TRUE.

3. Consider the online learning set-up where there exists a $f^* \in \mathcal{F}$ such that all the examples (x, y) satisfy $f^*(x) = y$. Let $|\mathcal{F}| = 5$ and number of examples is $m = 10$. Then, an upper bound on the number of mistakes the `consistent algorithm` makes in the worst-case in $m = 10$ rounds is 4. TRUE.

4. Consider a classification problem, where the input space is Euclidean. *Vipareeta Buddhi* claims that the kernelized k-NN classification with Gaussian kernel will be exactly same as k-NN classification with Euclidean distance. His claim is TRUE.

5. Consider the l_2 regularized logistic regression model for binary classification with linearly separable training data (linearly separable in the feature space). The corresponding ERM problem always has a (finite) solution. TRUE.

6. The number of clusters obtained with a 5-component Gaussian mixture model can be 6. FALSE.

7. If $x \in \mathbb{R}^r$ and $\phi(x)$ is the vector of all possible monomials involving entries of x upto degree d, then the dimensionality of $\phi(x)$ is $O(d^r)$. FALSE.

8. A kernel is a valid kernel iff its value is always non-negative/positive. FALSE.

9. There is no (unknown or to be tuned) regularization hyperparameter in case of the hard-margin SVM. TRUE.

10. There are examples of models that can be understood as parametric models, and can also be understood as non-parametric models. TRUE.

11. "*Complex concepts can be explained using appropriate examples*" – this is the philosophy behind all the formal machine learning set-ups you studied in this course. FALSE.

12. Consider the binary classification task under hinge-loss. Let $\mathcal{F}_1$ be the l_2 regularized linear model with margin atleast 1 and let $\mathcal{F}_2$ be the l_2 regularized linear model with margin atleast 2. Then, modelling error with $\mathcal{F}_1$ is $\leq$ that with $\mathcal{F}_2$. TRUE.

13. Consider the binary classification task under hinge-loss. Let $\mathcal{F}_1$ be the l_2 regularized linear model with margin atleast 1 and let $\mathcal{F}_2$ be the l_2 regularized linear model with margin atleast 2. Then, standard estimation error bound with $\mathcal{F}_1$ is $\leq$ that with $\mathcal{F}_2$. FALSE.

14. The definition of clustering we employed in our lecture is: grouping of inputs such that inputs within groups are similar to each other than inputs across different groups. FALSE.

15. Let $k_1 : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ and $k_2 : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be two valid kernels[1] over the domain, $\mathcal{X}$. Define a new function k by: $k(z) \equiv k_1(z)k_2(z)\ \forall\ z \in \mathcal{X} \times \mathcal{X}$. Then, k is a valid kernel over $\mathcal{X}$. TRUE.

16. Typically, models with lower estimation error tend to have higher approximation error and vice-versa. TRUE.

17. *Sukshma Buddhi* and *Sthula Buddhi* both plan to deploy a particular Binary classification model that has one hyperparameter, $\gamma \in (0, 1]$. Both of them use 10-fold CV for estimating γ. However, *Sthula Buddhi* considers the range of candidate values for γ as $\{1, \frac{1}{2}, \ldots, \frac{1}{2^9}\}$, whereas *Sukshma Buddhi*, being an expert in numerical optimization, considers entire interval $(0, 1]$ and finds the "best" (upto numerical errors[2]). Let us assume both have access to the same data and consider the same CV folds. Then, smaller the difference between the CV errors with the two models, less likely it is that *Sthula Buddhi*'s model will perform better, when deployed, than *Sukshma Buddhi*'s model. FALSE.

18. The backprop algorithm solves the stochastic optimization problem arising in case of neural-network based logistic regression. FALSE.

19. Linear models are not universal approximators, whereas kernel-based models are universal approximators. FALSE.

[1]Kernel here refers to kernels in SVM etc. and NOT smoothing kernels. Nor it is popcorn kernels ;) .
[2]Let's assume that the CV error happens to be a "nice" function to optimize numerically over $(0, 1]$.

20. Typical estimation error bounds for linear models asymptotically decay to zero as number of samples grows to infinity. However, the bounds are not independent of input-dimensionality. TRUE.

21. Typical estimation error bounds for l_2 regularized linear models are independent of input-dimensionality. TRUE.

22. In kernel-based models, the input-feature-map needs to be designed carefully. FALSE.

23. Consider two different 3-arm bandit problems:

    P1: $q(a_1) = 1, q(a_2) = 1 - \epsilon, q(a_3) = 9 + \epsilon$.

    P2: $q(a_1) = 1, q(a_2) = 9 - \epsilon, q(a_3) = 9 + \epsilon$.

    where, $\epsilon = 1e-6$. Then, as per the UCB algorithm analysis presented during the lecture, P1 is a simpler problem than P2. TRUE.

24. Applications of Gaussian mixture models go beyond clustering: for example, they have potential to be applied in regression tasks. TRUE.

25. The MLE problem arising in case of parameter estimation with Gaussian mixture models is popularly solved by Gradient Descent (or SGD). FALSE.

26. PCA can be understood as a set-up employing l_2 regularized linear models. Hence it is expected to be free from the curse of dimensionality. TRUE.

27. 1-class SVM can be employed for support estimation tasks as well as for clustering tasks. TRUE.

28. Training neural networks implicitly performs representation learning; however this representation learning can neither be categorized as supervised learning nor as unsupervised learning. TRUE.

29. PCA set-up is a special case of the autoassociative neural network set-up. TRUE.

30. In neural network modelling, the Bayes optimal corresponding to the underlying (unknown) joint likelihood function in the supervised learning set-up is modelled directly. FALSE.

31. In kernelized PCA, the top few eigenvectors of the gram matrix are computed instead of those of the sample correlation/covariance matrix. TRUE.

32. The only difference between online learning and batch learning is that in the former set-up the training samples arrive sequentially and cannot be revisited. FALSE.

33. The UCB algorithm performs both exploration as well as exploitation; whereas the soft-max algorithm only performs exploration. FALSE.

34. The Gaussian model is not a well-suited model for clustering tasks. TRUE.

(2) Simple Exercises

① Backprop computes gradient of terms in objective,

We will show this on the first term: $\left(v_1\sigma(w_1) + v_2\sigma(w_2) - 1\right)^2 \equiv g(v_1, v_2, w_1, w_2)$

$$\frac{\partial g}{\partial v_1} = 2\left(v_1\sigma(w_1) + v_2\sigma(w_2) - 1\right)\left(\sigma(w_1)\right)$$

$$\frac{\partial g}{\partial v_2} = 2\left(v_1\sigma(w_1) + v_2\sigma(w_2) - 1\right)\left(\sigma(w_2)\right)$$

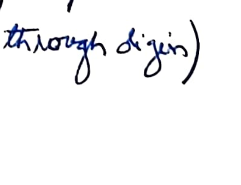
$$\frac{\partial g}{\partial w_1} = \begin{cases} 2\left(v_1\sigma(w_1) + v_2\sigma(w_2) - 1\right)v_1 & \text{if } w_1 > 0 \\ 0 & \text{else} \end{cases}$$

though not differentiable at exact $w_1 = 0 / w_2 = 0$. Formally, notion of sub-gradients is used.

$$\frac{\partial g}{\partial w_2} = \begin{cases} 2\left(v_1\sigma(w_1) + v_2\sigma(w_2) - 1\right)v_2 & \text{if } w_2 > 0 \\ 0 & \text{else} \end{cases}$$

(2) $\longrightarrow$ Solution for first blank $\rightarrow$ "Does Not Exist".

Method ① $k(x,y) = x^T y \longrightarrow \varphi(x) = x$ is a feature map.



$\downarrow$

But data is Not linearly separable (line through origin)

$\therefore$ No Solution

Method 2 ERM is

$$\min_{\alpha_1, \alpha_2 \in \mathbb{R}} \frac{1}{2} [\alpha_1, \alpha_2] \begin{bmatrix} 2 & 4 \\ 4 & 8 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$$

s.t. $\alpha_1 2 + \alpha_2 4 \geq 1$

$-1(\alpha_1 4 + \alpha_2 8) \geq 1$

$\longrightarrow$ Not feasible!

→ Solution for next box (3 marks)

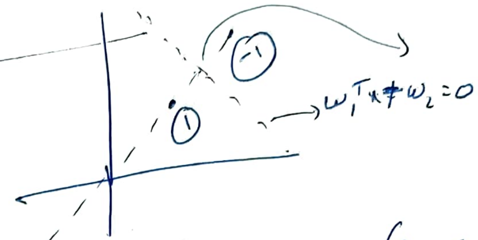
Method 1

$$k(x,y) = 1 + x^T y \longrightarrow \varphi(x) = \begin{bmatrix} x \\ 1 \end{bmatrix} \text{ is a feature map.}$$

$$\downarrow$$

$$w^T \varphi(x) = w_1^T x + w_2 \longrightarrow \text{lines need not pass through origin.}$$

SVM solution will be ~~to~~ perpendicular bisector as explained in lecture.



$\therefore$ $w_1^* = \begin{bmatrix} \gamma \\ \gamma \end{bmatrix}$ for some γ → variable (reparametrize) & ~~w_2~~
at optimality

→ Parallel to line joining $\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}2\\2\end{bmatrix}$

→ $w_1^T x + w_2 = 0$ passes through mid point of $\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}2\\2\end{bmatrix}$

$\therefore$ $\gamma 1.5 + 1.5\gamma + w_2 = 0 \Rightarrow w_2 = -3\gamma$

$\therefore$ From geometry we know the optimal ~~classifier discriminator~~ hyperplane is

$$f(x) = \gamma x_1 + \gamma x_2 - 3\gamma = 0$$

Use ERM for finding γ.

$$\min_{\gamma} \frac{1}{2}(\gamma^2 + \gamma^2 + 9\gamma^2) \qquad = \qquad \min_{\gamma} \frac{11}{2}\gamma^2 \qquad \Rightarrow \gamma^* = -1$$

$$\text{s.t.} \quad 1(\gamma + \gamma - 3\gamma) \geq 1 \qquad\qquad \text{s.t.} \quad \gamma \leq -1$$

$$-1(2\gamma + 2\gamma - 3\gamma) \geq 1$$

$\therefore$ $f(x) = -x_1 - x_2 + 3 = 0$ is optimal hyperplane

$$f\left(\begin{bmatrix}2\\3\end{bmatrix}\right) = -2-3+3 = -2.$$

Method 2

ERM is

$$\min_{\alpha_1, \alpha_2 \in \mathbb{R}} \frac{1}{2}\begin{bmatrix}\alpha_1 & \alpha_2\end{bmatrix}\begin{bmatrix}3 & 5\\ 5 & 9\end{bmatrix}\begin{bmatrix}\alpha_1\\ \alpha_2\end{bmatrix}$$

s.t.

$$1(\alpha_1 3 + 5\alpha_2) \geq 1$$

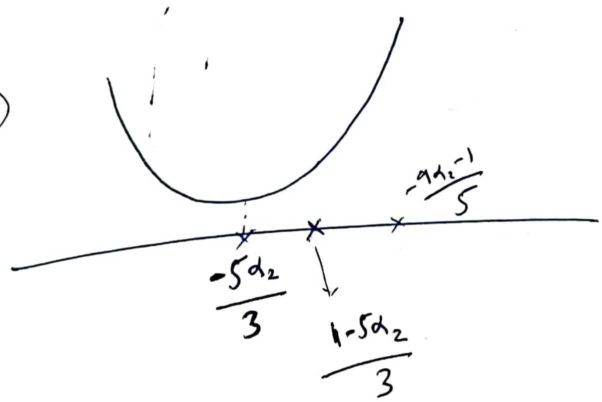
$$-1(\alpha_1 5 + \alpha_2 9) \geq 1$$

Solve wrt α_1 first (assume α_2 is unknown constant)

$\longrightarrow$

$$\min_{\alpha_1 \in \mathbb{R}} 3\alpha_1^2 + 10\alpha_1\alpha_2 + 9\alpha_2^2$$

s.t. $\frac{1-5\alpha_2}{3} \leq \alpha_1 \leq \frac{-9\alpha_2-1}{5}$



$\therefore$ Minimum achieved at $\alpha_1^* = \frac{1-5\alpha_2}{3}$.

Eliminating α_1 from ERM:

$$\min_{\alpha_2 \in \mathbb{R}} 3\left(\frac{1-5\alpha_2}{3}\right)^2 + 10\left(\frac{1-5\alpha_2}{3}\right)\alpha_2 + 9\alpha_2^2$$

s.t. $\frac{1-5\alpha_2}{3} \leq \frac{-9\alpha_2-1}{5}$

$$= \min_{\alpha_2 \in \mathbb{R}} \quad 2\alpha_2^2 + 1 \longrightarrow \text{solution } \alpha_2^* = -4.$$

$$\underline{\underline{\text{s.t.}}} \quad \alpha_2 \leq -4$$

$$\Downarrow$$

$$\alpha_1^* = \frac{1 - 5(-4)}{3} = 7$$

$\therefore$ optimal prediction function is

$$f(x) = 7 k_2\left(\begin{bmatrix}1\\1\end{bmatrix}, x\right) - 4 k_2\left(\begin{bmatrix}2\\2\end{bmatrix}, x\right)$$

$$f\left(\begin{bmatrix}2\\3\end{bmatrix}\right) = 7(6) - 4(11) = -2$$

3) True or False

~~① 10 is an upper bound (covered in lecture)~~
~~Additionally, note, unlike in lecture, |H| is finite~~
~~So another upper bound is possible.~~

① Covered in lecture $\otimes$ $\log(|\mathcal{F}|) = \log(8) = 3.$

② No. clusters is 5 for 5-GMM $\Rightarrow$ threshold is high enough.

$\Downarrow$

clusters are elliptical
(level sets of Gaussians are elliptical)

③

(3) 10 is the trivial upper bound from lecture.

Here, unlike in lecture, $|\mathcal{F}| < \infty$ so we get another upper bound.

Note that consistent algo atleast removes one function

has mistake (the one picked while committing mistake)

$\therefore$ $\#$mistakes $\leq |\mathcal{F}| - 1$ (-1 because $f^* \in \mathcal{F}$) last one $\rightarrow$

$= 4$

(4) kNN with Gaussian kernel uses distance as

$$\sqrt{k(x,x) + k(y,y) - 2k(x,y)}$$

See section 5.11.1 in Summary Notes

$$= \sqrt{e^{-\frac{1}{2\sigma^2}\|x-x\|^2} + e^{-\frac{1}{2\sigma^2}\|y-y\|^2} - 2e^{-\frac{1}{2\sigma^2}\|x-y\|^2}}$$

$$= \sqrt{2 - 2e^{-\frac{1}{2\sigma^2}\|x-y\|^2}}$$

$$\|x-y\| \leq \|x-z\|$$

$$\iff \sqrt{2-2e^{-\frac{1}{2\sigma^2}\|x-y\|^2}} \leq \sqrt{2-2e^{-\frac{1}{2\sigma^2}\|x-z\|^2}}$$

$\therefore$ kNN will give same answer.

(5) If $\|w\|$ is not there in minimization, $w \rightarrow \infty$ but this would not happen.

(6) No. $\rightarrow$ no. of clusters will be ≤ 5

$\leq$ no. of peaks in pdf $= 5$

or $\rightarrow$ joined.

(7) $O(n^d)$ → See Pg 24 in summary notes first para.

(8)

$$G = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \rightarrow \geq 0$$

(9) YES → $B = 0$ (equations (5.17), (5.18) in Summary Notes)

(10) Eg. SVM

(11) Eg. Reinforcement learning is <u>Not</u> example-based.

(12) $\mathcal{F}_1 \supseteq \mathcal{F}_2 \Rightarrow$ model error $\mathcal{F}_1 \leq$ model error $\mathcal{F}_2$

(13) $\cancel{W_1}$ $\cancel{\frac{2}{\|w\|}} \leq$ ~~margin~~ $\frac{2}{\|w\|} \geq \frac{2}{W} \equiv$ → least margin in model.

$\mathcal{F}_1$ → $\cancel{W_1}$ $W_1 = 2$

$\mathcal{F}_2$ → $W_2 = 1$ → better.

Est error $\leq \sqrt{\frac{WK}{m}}$ ∴ better.

(14) See section 2.5.1 in summary notes.

(15) See theorem 3.1.2 in summary notes.

(16) See ~~section~~ chapter 6

(17) Analogical to Big Model vs small Model with same training error. Last para Pg 64 in Summary Notes.

(18) It is an algorithm to compute gradient of terms in objective. Not to solve an optimization problem.

(19) Both are universal approximators.

(20) Second para Pg 62 in Summary Notes.

(21) Section 3.1.2.1 in Summary Notes.

(22) Design kernel, which implicitly gives feature map.

(23) Second last para Pg 39 in Summary Notes.

(24) GMMs are universal approximators for likelihood functions.

(25) EM algorithm is popular.

(26) Problem 1 in Section 1 in this Paper.

(27) Sections 5.18, 5.18.1 in Summary Notes

(28) First para Pg 16

(29) Autoencoders with 1 hidden & ~~linear~~ identity activations

(30) Human neural processing is modelled.

(31) Section 5.17.1 in Summary Notes.

(32) Main difference is evaluation is online (after every example)

(33) Softmax also explore-exploit.

(34) Only one ~~peak~~ peak ⇒ Only one cluster.