Final Pset

Saturday, March 23, 2019 8:38 PM

1. The polynomial transform of order $Q = 10$ applied to $\mathcal{X}$ of dimens
sults in a $\mathcal{Z}$ space of what dimensionality (not counting the constan
$x_0 = 1$ or $z_0 = 1$)?

[a] 12

[b] 20

[c] 35

[d] 100

[e] None of the above

55. I thought we were approximating...

The feature space X has two dimensions. For a 10th order polynomial, not including constants, we have the following ter
We can express the "activation" of a z vector as follows

$$\sum_{i=1}^{10}\sum_{j=0}^{i} a_{ij} x_1^i x_2^{j-i}$$

$$\sum_{i=1}^{10}(i+1) = 10 + \sum_{i=1}^{10} i = 10 + 11 * \frac{10}{2} = 65$$

10+9+8+7+6+5+4+3+2+1 = 55

2. Recall that the average hypothesis $\bar{g}$ was based on tr
on different data sets $\mathcal{D}$ to get $g^{(\mathcal{D})} \in \mathcal{H}$, and taking
w.r.t. $\mathcal{D}$ to get $\bar{g}$. Which of the following models $\mathcal{H}$ c

~~**[a]** A singleton $\mathcal{H}$ ($\mathcal{H}$ has one hypothesis)~~

~~**[b]** $\mathcal{H}$ is the set of all constant, real-valued hypothe~~

~~**[c]** $\mathcal{H}$ is the linear regression model~~

[d] $\mathcal{H}$ is the logistic regression model

[e] None of the above

ion $d = 2$ re-
nt coordinate

ms:

raining the same model $\mathcal{H}$
the expected value of $g^{(\mathcal{D})}$
could result in $\bar{g} \notin \mathcal{H}$?

ses

[d] $\mathcal{H}$ is the logistic regression model

[e] None of the above

All hypothesis spaces are "convex spaces"
The "average" of any set constants is a constant.
The average of a set of linear models can be arranged into another linear model where the weight vector is the aver
commutative and associative properties repeatedly)
For the log-reg models, this is not the case.

- **Overfitting**

3. Which of the following statements is *false*?

[a] If there is overfitting, there must be two or more hypotheses that different values of E_{in}. True! Overfitting is part of learning, which is discriminating hypo

[b] If there is overfitting, there must be two or more hypotheses that different values of E_{out}. True! Again, we are learning in a way that degrades Eout

[c] If there is overfitting, there must be two or more hypotheses that different values of $(E_{\text{out}} - E_{\text{in}})$. Eout2 > Eout1. Ein2 < Ein1. Eo
we can re-arrange to put c1 a

[d] We can always determine if there is overfitting by comparing the valu $(E_{\text{out}} - E_{\text{in}})$. Sure, but if we had that so readily available.....

[e] We cannot determine overfitting based on one hypothesis only.
true

Overfitting in general, is the specific phenomen of "learning" in a way that we decrease Ein, but are increasing Eout.
problem/function we are trying to learn about.

4. Which of the following statements is true?

[a] Deterministic noise cannot occur with stochastic noise.

[b] Deterministic noise does not depend on the hypothesis s

[c] Deterministic noise does not depend on the target functi

[d] Stochastic noise does not depend on the hypothesis set.

[e] Stochastic noise does not depend on the target distribut

age weight vector of the linear model (leverage

have

theses.

have

have

ut2 - Ein2 ≈ Eout1 + c1 - Ein1 + c2. c1 and c2 are non-negative, therefore
nd c2 on one side

es of

We do this by "learning the dataset" and not the

et.

Determinsitic noise is another word for bias. Bias is the inability of a Hset to represent the target function.

ion.

ion.

Don't know what this means.

• Regularization

5. The regularized weight $\mathbf{w}_{\text{reg}}$ is a solution to:

$$\text{minimize} \quad \frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n - y_n)^2 \quad \text{subject to} \quad \mathbf{w}^{\mathrm{T}}\Gamma^{\mathrm{T}}$$

where Γ is a matrix. If $\mathbf{w}_{\text{lin}}^{\mathrm{T}}\Gamma^{\mathrm{T}}\Gamma\mathbf{w}_{\text{lin}} \leq C$, where $\mathbf{w}_{\text{lin}}$ is t solution, then what is $\mathbf{w}_{\text{reg}}$?

[a] $\mathbf{w}_{\text{reg}} = \mathbf{w}_{\text{lin}}$

[b] $\mathbf{w}_{\text{reg}} = \Gamma\mathbf{w}_{\text{lin}}$

[c] $\mathbf{w}_{\text{reg}} = \Gamma^{\mathrm{T}}\Gamma\mathbf{w}_{\text{lin}}$

[d] $\mathbf{w}_{\text{reg}} = C\Gamma\mathbf{w}_{\text{lin}}$

[e] $\mathbf{w}_{\text{reg}} = C\mathbf{w}_{\text{lin}}$

If wlin is vanilla lin reg solution, and it meets our regulari

Then we done! Our "simple" / "reduced" hypothesis set

Turns out, our concerns over stochastic noise were great

6. Soft-order constraints that regularize polynomial models can be

[a] written as hard-order constraints

[b] translated into augmented error

[c] determined from the value of the VC dimension

[d] used to decrease both E_{in} and E_{out} More constraint, Ein is non-decreasing!

[e] None of the above is true

• Bayesian Priors

19. Let $f \in [0, 1]$ be the unknown probability of getting a heart attack in a certain population. Notice that f is just a constant, not a fu

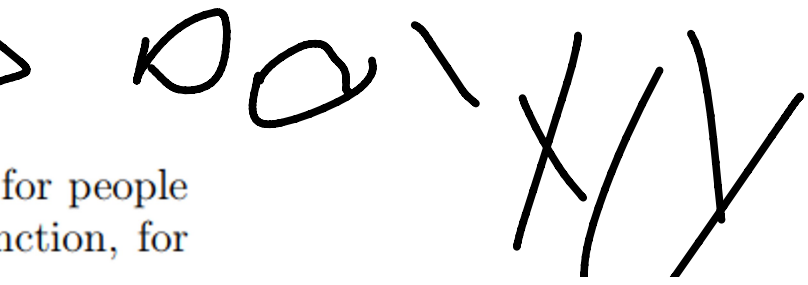
$^{\mathrm{T}}\Gamma\mathbf{w} \leq C,$

the linear regression

ization constriant….

still GOT THE JOB DONE.

tly exagerated :)



for people
nction, for

19. Let $f \in [0, 1]$ be the unknown probability of getting a heart attack in a certain population. Notice that f is just a constant, not a fu simplicity. We want to model f using a hypothesis $h \in [0, 1]$. Before data, we assume that $P(h = f)$ is uniform over $h \in [0, 1]$ (the prior) one person from the population, and it turns out that he or she ha attack. Which of the following is true about the posterior probability t given this sample point?

[a] The posterior is uniform over $[0, 1]$.

[b] The posterior increases linearly over $[0, 1]$.

[c] The posterior increases nonlinearly over $[0, 1]$.

[d] The posterior is a delta function at 1 (implying f has to be 1).

[e] The posterior cannot be evaluated based on the given informati

Probability of 0 i

Before we had a flat thing.
Now, we learn on |D| = 1. Specifically, we know P(h=f|D) = P uniform and P(DO is constant throughout, we can reformat t simply compute then "normalize" via integration. The shape potential "constant value" of f is linear asape.)

- **Aggregation**

20. Given two learned hypotheses g_1 and g_2, we construct the aggregate hypothesis g given by $g(\mathbf{x}) = \frac{1}{2}(g_1(\mathbf{x}) + g_2(\mathbf{x}))$ for all $\mathbf{x} \in \mathcal{X}$. If we use mean-squared error, which of the following statements is true?

[a] $E_{\text{out}}(g)$ cannot be worse than $E_{\text{out}}(g_1)$.

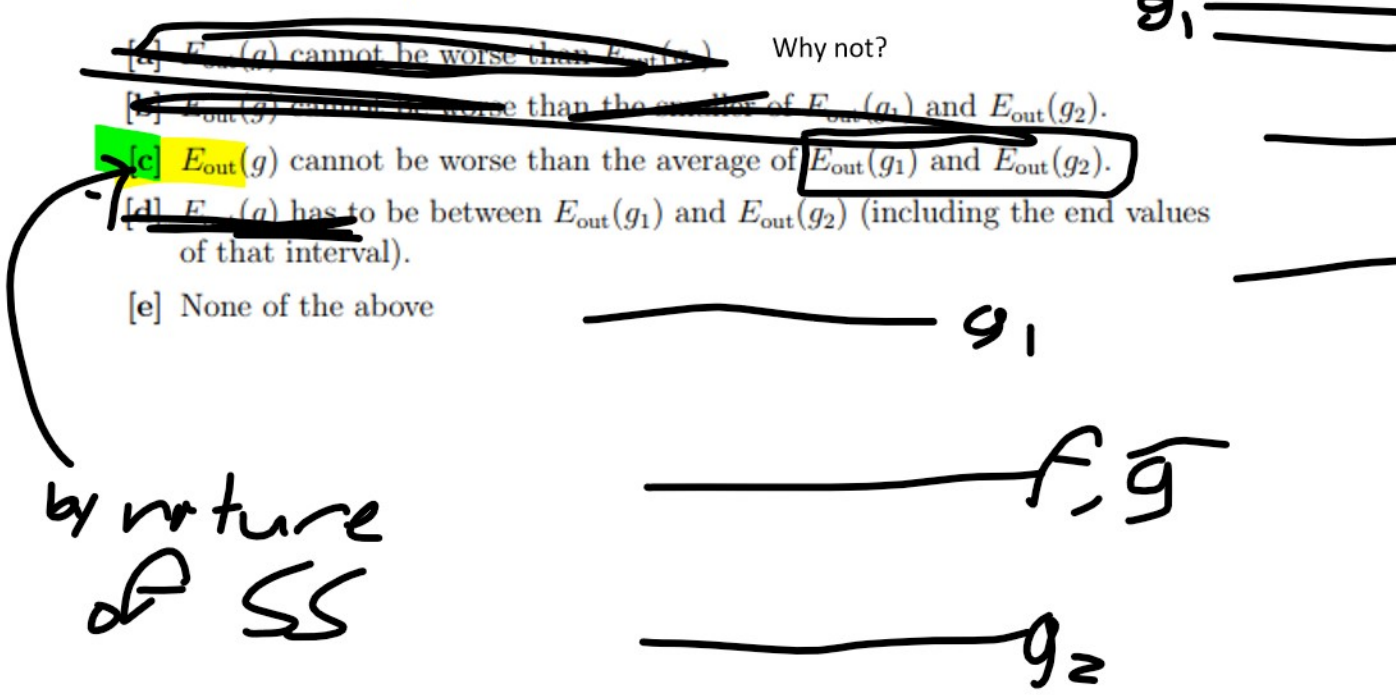
Why not?

[b] $E_{\text{out}}(g)$ cannot be worse than the smaller of $E_{\text{out}}(g_1)$ and $E_{\text{out}}(g_2)$.

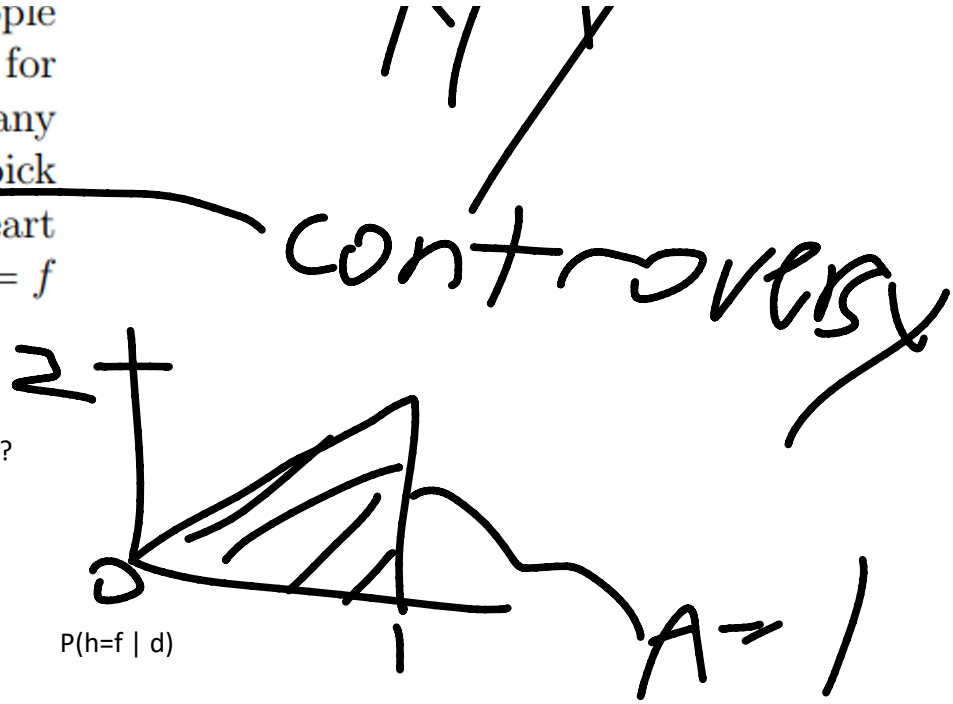
[c] $E_{\text{out}}(g)$ cannot be worse than the average of $E_{\text{out}}(g_1)$ and $E_{\text{out}}(g_2)$.

[d] $E_{\text{out}}(g)$ has to be between $E_{\text{out}}(g_1)$ and $E_{\text{out}}(g_2)$ (including the end values of that interval).

[e] None of the above



ıor people
nction, for
we see any
. We pick
ad a heart
that $h = f$



is 0 tho rite?

P(h=f | d)

ion.

P(D|h=f)*P(h = f) / P(D). Given our prior is
o P(h=f|D) = alpha P(D|h=f), which, we
of this "posterior distribution" of the

7. Set $\lambda = 1$ and do not apply a feature transform (i.e., use $\mathbf{z} = \mathbf{x} = (1, x_1, x$
Which among the following classifiers has the lowest E_{in}?

[a] 5 versus all

[b] 6 versus all

[c] 7 versus all

[d] 8 versus all

[e] 9 versus all

8. Now, apply a feature transform $\mathbf{z} = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$, and set $\lambda =$
Which among the following classifiers has the lowest E_{out}?

[a] 0 versus all

[b] 1 versus all

[c] 2 versus all

[d] 3 versus all

[e] 4 versus all

9. If we compare using the transform versus not using it, and apply that to '0 versus all' through '9 versus all', which of the following statements is correct for $\lambda = 1$?

[a] Overfitting always occurs when we use the transform. False, some improve

[b] The transform always improves the out-of-sample performance by at least 5% (E_{out} with transform $\leq 0.95 E_{\text{out}}$ without transform). False, some imr

[c] The transform does not make any difference in the out-of-sample performance.

Definitively untrue

$$\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{z}_n - y_n\right)^2 \;+\; \frac{\lambda}{N}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

1.

$$\begin{aligned} E_{\mathrm{in}}(\mathbf{w}) &= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n - y_n\right)^2 \\ &= \frac{1}{N}\|\mathrm{X}\mathbf{w} - \mathbf{y}\|^2 \\ &= \frac{1}{N}(\mathbf{w}^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{X}\mathbf{w} - 2\mathbf{w}^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y}), \end{aligned}$$

$$\nabla E_{\mathrm{in}}(\mathbf{w}) = \frac{2}{N}(\mathrm{X}^{\mathrm{T}}\mathrm{X}\mathbf{w} - \mathrm{X}^{\mathrm{T}}\mathbf{y}).$$ +2\lambda/N w^T = 0

bove some dont

(3.3)

(3.4)

[d] The transform always worsens the out-of-sample performance by at least 5%.

[e] The transform improves the out-of-sample performance of '5 versus all,' but by less than 5%.

10. Train the '1 versus 5' classifier with $\mathbf{z} = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$ with $\lambda = 0.01$ and $\lambda = 1$. Which of the following statements is correct?

[a] Overfitting occurs (from $\lambda = 1$ to $\lambda = 0.01$).

[b] The two classifiers have the same E_{in}.

[c] The two classifiers have the same E_{out}.

[d] When λ goes up, both E_{in} and E_{out} go up.

[e] When λ goes up, both E_{in} and E_{out} go down.

- **Support Vector Machines**

11. Consider the following training set generated from a target function $f : \mathcal{X} \to \{-1, +1\}$ where $\mathcal{X} = \mathbb{R}^2$

$$\begin{array}{ccc} \mathbf{x}_1 = (1,0), y_1 = -1 & \mathbf{x}_2 = (0,1), y_2 = -1 & \mathbf{x}_3 = (0,-1), y_3 = -1 \\ \mathbf{x}_4 = (-1,0), y_4 = +1 & \mathbf{x}_5 = (0,2), y_5 = +1 & \mathbf{x}_6 = (0,-2), y_6 = +1 \\ & \mathbf{x}_7 = (-2,0), y_7 = +1 & \end{array}$$

Transform this training set into another two-dimensional space $\mathcal{Z}$

$$z_1 = x_2^2 - 2x_1 - 1 \qquad z_2 = x_1^2 - 2x_2 + 1$$

Using geometry (not quadratic programming), what values of $\mathbf{w}$ (without w_0) and b specify the separating plane $\mathbf{w}^{\mathrm{T}}\mathbf{z} + b = 0$ that maximizes the margin in the $\mathcal{Z}$ space? The values of w_1, w_2, b are:

[a] $-1,\ 1,\ -0.5$

[b] $1,\ -1,\ -0.5$

[c] $1,\ 0,\ -0.5$

[d] $0,\ 1,\ -0.5$

[e] None of the above would work.

12. Consider the same training set of the previous problem, but instead of explicitly

12. Consider the same training set of the previous problem, but instead of explicitly transforming the input space $\mathcal{X}$, apply the hard-margin SVM algorithm with the kernel

$$K(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^{\mathrm{T}}\mathbf{x}')^2$$

(which corresponds to a second-order polynomial transformation). Set up the expression for $\mathcal{L}(\alpha_1 ... \alpha_7)$ and solve for the optimal $\alpha_1, ..., \alpha_7$ (numerically, using a quadratic programming package). The number of support vectors you get is in what range?

[a] 0-1

[b] 2-3

[c] 4-5

[d] 6-7

[e] >7

- **Radial Basis Functions**

We experiment with the RBF model, both in regular form (Lloyd + pseudo-inverse) with K centers:

$$\text{sign}\left(\sum_{k=1}^{K} w_k \exp\left(-\gamma \, ||\mathbf{x} - \mu_k||^2\right) + b\right)$$

(notice that there is a bias term), and in kernel form (using the RBF kernel in hard-margin SVM):

$$\text{sign}\left(\sum_{\alpha_n > 0} \alpha_n y_n \exp\left(-\gamma \, ||\mathbf{x} - \mathbf{x}_n||^2\right) + b\right).$$

The input space is $\mathcal{X} = [-1, 1] \times [-1, 1]$ with uniform probability distribution, and the target is

$$f(\mathbf{x}) = \text{sign}(x_2 - x_1 + 0.25 \sin(\pi x_1))$$

which is slightly nonlinear in the $\mathcal{X}$ space. In each run, generate 100 training points at random using this target, and apply both forms of RBF to these training points. Here are some guidelines:

- Repeat the experiment for as many runs as needed to get the answer to be stable (statistically away from flipping to the closest competing answer).
- In case a data set is not separable in the '$\mathcal{Z}$ space' by the RBF kernel using hard-margin SVM, discard the run but keep track of how often this happens, if ever.

- When you use Lloyd's algorithm, initialize the centers to random points in $\mathcal{X}$ and iterate until there is no change from iteration to iteration. If a cluster becomes empty, discard the run and repeat.

13. For $\gamma = 1.5$, how often do you get a data set that is not separable by the RBF kernel (using hard-margin SVM)? *Hint: Run the hard-margin SVM, then check that the solution has* $E_{\text{in}} = 0$.

[a] $\leq 5\%$ of the time

[b] $> 5\%$ but $\leq 10\%$ of the time

[c] $> 10\%$ but $\leq 20\%$ of the time

[d] $> 20\%$ but $\leq 40\%$ of the time

[e] $> 40\%$ of the time

F
k

14. If we use $K = 9$ for regular RBF and take $\gamma = 1.5$, how often does the kernel form beat the regular form (excluding runs mentioned in Problem 13 and runs with empty clusters, if any) in terms of E_{out}?

[a] $\leq 15\%$ of the time

[b] $> 15\%$ but $\leq 30\%$ of the time

[c] $> 30\%$ but $\leq 50\%$ of the time

[d] $> 50\%$ but $\leq 75\%$ of the time

[e] $> 75\%$ of the time

15. If we use $K = 12$ for regular RBF and take $\gamma = 1.5$, how often does the kerne form beat the regular form (excluding runs mentioned in Problem 13 and run with empty clusters, if any) in terms of E_{out}?

[a] $\leq 10\%$ of the time

[b] $> 10\%$ but $\leq 30\%$ of the time

[c] $> 30\%$ but $\leq 60\%$ of the time

[d] $> 60\%$ but $\leq 90\%$ of the time

[e] $> 90\%$ of the time

16. Now we focus on regular RBF only, with $\gamma = 1.5$. If we go from $K = 9$ cluster to $K = 12$ clusters (only 9 and 12), which of the following 5 cases happens mos often in your runs (excluding runs with empty clusters, if any)? Up or dow means strictly so.

[a] E_{in} goes down, but E_{out} goes up.

[b] E_{in} goes up, but E_{out} goes down.

[c] Both E_{in} and E_{out} go up.

[d] Both E_{in} and E_{out} go down.

[e] E_{in} and E_{out} remain the same.

17. For regular RBF with $K = 9$, if we go from $\gamma = 1.5$ to $\gamma = 2$ (only 1.5 and 2 which of the following 5 cases happens most often in your runs (excluding ru with empty clusters, if any)? Up or down means strictly so.

[a] E_{in} goes down, but E_{out} goes up.

[b] E_{in} goes up, but E_{out} goes down.

[c] Both E_{in} and E_{out} go up.

[d] Both E_{in} and E_{out} go down.

[e] E_{in} and E_{out} remain the same.

s
t
n

),
ns

18. What is the percentage of time that regular RBF achieves $E_{\text{in}} = 0$ with $K = 9$ and $\gamma = 1.5$ (excluding runs with empty clusters, if any)?

[a] $\leq 10\%$ of the time

[b] $> 10\%$ but $\leq 20\%$ of the time

[c] $> 20\%$ but $\leq 30\%$ of the time

[d] $> 30\%$ but $\leq 50\%$ of the time

[e] $> 50\%$ of the time

Hard-Margin SVM with Kernel

1: Construct $\mathrm{Q_D}$ from the kernel K, and $\mathrm{A_D}$:

$$\mathrm{Q_D} = \begin{bmatrix} y_1 y_1 \mathrm{K}_{11} & \cdots & y_1 y_N \mathrm{K}_{1N} \\ y_2 y_1 \mathrm{K}_{21} & \cdots & y_2 y_N \mathrm{K}_{2N} \\ \vdots & \vdots & \vdots \\ y_N y_1 \mathrm{K}_{N1} & \cdots & y_N y_N \mathrm{K}_{NN} \end{bmatrix} \quad \text{and} \quad \mathrm{A_D} = \begin{bmatrix} \mathbf{y}^{\mathrm{T}} \\ -\mathbf{y}^{\mathrm{T}} \\ \mathrm{I}_{N \times N} \end{bmatrix},$$

where $\mathrm{K}_{mn} = K(\mathbf{x}_m, \mathbf{x}_n)$. (K is called the *Gram* matrix.)

2: Use a QP-solver to optimize the dual problem:

$$\boldsymbol{\alpha}^* \leftarrow \mathsf{QP}(\mathrm{Q_D}, -\mathbf{1}_N, \mathrm{A_D}, \mathbf{0}_{N+2}).$$

3: Let s be any support vector for which $\alpha_s^* > 0$. Compute

$$b^* = y_s - \sum_{\alpha_n^* > 0} y_n \alpha_n^* K(\mathbf{x}_n, \mathbf{x}_s).$$

4: Return the final hypothesis

$$g(\mathbf{x}) = \operatorname{sign}\left(\sum_{\alpha_n^* > 0} y_n \alpha_n^* K(\mathbf{x}_n, \mathbf{x}) + b^* \right).$$

9

x