Eli Pinkus
CS 156a
Final

Problem 1

We know that a polynomial transform of order $Q = 10$ will contain all polynomials of x_1, x_2 between orders 1 and $Q = 10$. We know that for some order q, we have $q + 1$ such elements. See these examples:

$$q = 1:$$
$$\vec{z} = (x_1, x_2)$$

$$q = 2:$$
$$\vec{z} = (x_1^2, x_1 x_2, x_2^2)$$

So we can express the dimensionality of the $\mathcal{Z}$ space for a Q dimensional polynomial transform as:

$$\sum_{q=1}^{Q} (q + 1)$$

So for $Q = 10$:

$$\sum_{q=1}^{10} (q + 1) = 65$$

So the answer is E

Problem 2

A

With a singleton hypothesis set we know that all g are equal to the singleton hypothesis. Thus, $\bar{g} \in \mathcal{H}$ will also be that singleton hypothesis

B

The average of any real values is also a real value thus $\bar{g} \in \mathcal{H}$

C

Linear regression implements a line/hyperplane that is defined linearly by $\vec{w}$ which is a vector of real values. Thus, the average hypothesis, $\bar{g}$ will also be defined by a vector of real values which means that in this case, $\bar{g} \in \mathcal{H}$.

D

Logistic regression implements a nonlinear surface that is defined by $\vec{w}$ (and a non-linear function), thus the average hypothesis, which would involve a nonlinear combination of $\vec{w}^T x$'s could be a hypothesis that is not in the hypothesis set of the logistic regression model:

Thus, the answer is D

Problem 3

We know that overfitting is marked by a decrease in E_{in} while E_{out} increases as a model changes from one to the next. In order to observe this one would need to compare at least two models and look for a decrease in E_{in} while E_{out} increases. Thus, we would need a changing (and decreasing) E_{in} and a changing (and increasing) E_{out} and therefore we would also need to have a changing (and increasing) $(E_{\text{out}} - E_{\text{in}})$. Thus A,B,C,E are true statements.

It is not enough to compare values of $(E_{\text{out}} - E_{\text{in}})$ because if you only examined this quantity it would be impossible to determine whether a change in the quantity was due to a decreasing E_{in} *and* an increasing E_{out} as is necessary to indentify overfitting. $(E_{\text{out}} - E_{\text{in}})$ could increase due to a slightly decreasing E_{out} paired with a more substantial decreasing of E_{in}.

The answer is D

Problem 4

Deterministic noise is noise brought about by a model's inability to fit the target function whereas stochastic noise is random noise.

A

Stochastic noise is considered inherent to the target, while deterministic noise comes from the model used to try and approximate the target, thus they can occur simultaneously

B

Deterministic noise definitely depends on the hypothesis set because it is noise from the inability of the hypothesis set to fit the target complexity.

C

For the same reason as B, deterministic noise definitely depends on the target function

E

Stochastic noise is random noise that is the result of a noisy target function. If we account for that noise with a target distribution, we have that $P(x|y) = f(x) + \text{stochastic noise}$ and the stochastic noise is described by the properties of the target distribution

D

Stochastic noise is a property of the target distribution. As such it doesn't matter what model is used in an attempt to predict the target as the noise is considered to be inherent to the situation that one is trying to model.

Thus, D is true and the answer is D

Problem 5

We have the solution to the regularized linear regression w_{reg}, and we are asked to minimize linear regression with a constraint. We are also told that the regression solution obeys the constraint, thus it would also be the regularized solution. We know that $\mathcal{H}_{\text{reg}} \subset \mathcal{H}_{\text{lin}}$, so if $g_{\text{lin}} \in \mathcal{H}_{\text{reg}}$, then we know $g_{\text{lin}} = g_{\text{reg}} \Rightarrow w_{\text{lin}} = w_{\text{reg}}$ so the answer is A

Problem 6

A

Hard constraint requires that certain w_q take on a particular value, whereas in soft constrained w's can range with a view to reducing E_{in} but they can only range as far as they collectively obey the regularization constraint. For this reason you can't write a soft order constraint as a hard order constraint.

C

While introducing a constraint can reduce the VC dimension, the VC dimension does not contain the necessary information to recover what constraint is being used since the relationship is not one to one.

D

Adding a regularization constraint makes fitting the in-sample data more difficult which means that E_{in} will not decrease with the addition of a soft order constraint.

B

$$E_{\text{aug}}(h) = E_{\text{in}}(h) + \frac{\lambda}{N}\Omega(h)$$

where Ω is a regularizer. So with a soft order constraint we have something like:

$$E_{\text{aug}}(w) = E_{\text{in}}(w) + \frac{\lambda}{N}w^T w$$

so we can definitely translate it into an augmented error and the answer is B

Problem 7

The simulation returned that $k = 8$ gave the lowest E_{in} the answer is D

Problem 8

The simulation returned that $k = 1$ gave the lowest E_{out} the answer is B

Problem 9

A False

In some case E_{out} goes down when the transform is applied so that indicates that there isn't always overfitting

B False
There are a few cases where E_{out} does not change

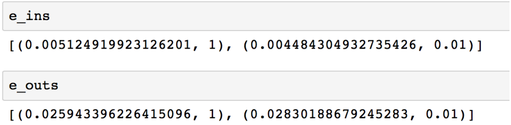
C False
There are many cases where E_{out} changes

D False
There are cases where E_{out} improves with transform

E True
With transform 5 vs all $E_{\text{out}} = 0.07922$
Without transform 5 vs all $E_{\text{out}} = 0.07972$

Which is a change of $< 5\%$
So the answer is E

Problem 10



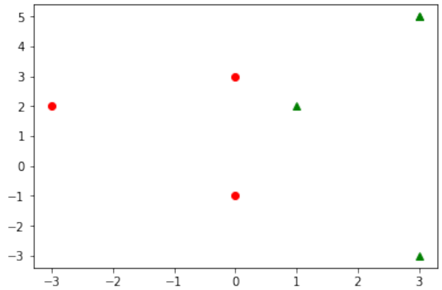
```
e_ins
[(0.005124919923126201, 1), (0.004484304932735426, 0.01)]

e_outs
[(0.025943396226415096, 1), (0.02830188679245283, 0.01)]
```

The above is the simulation output in the form (Error, associated λ)
By this we can see that B,C,D,E are false

We can also see that from $\lambda = 1$ to $\lambda = 0.01$, E_{in} goes down while E_{out} goes up which indicates that there is overfitting so the answer is A

Problem 11

We identify the “support vectors” by inspection of the $\mathcal{Z}$ space plot:



We can easily define the separating plane that maximized the margin in the $\mathcal{Z}$ space:

We have the dividing line $z_1 = \frac{1}{2}$, So we have that the classification according to this boundry depends only on z_1 so we have $w_2 = 0$

So we have the dividing line:

$$w_1 z_1 + b = \frac{w_1}{2} + b = 0 \Rightarrow w_1 = -2b$$

$1,0,-0.5$ is the only option that satisfies these conditions thus the answer is C

Problem 12

The number of support vectors is 5 so the answer is C

Problem 13

The SVM returned $E_{\text{in}} \neq 0$, 0 times in 1000 runs so the answer is A

Problem 14

The SVM beat the regular RBF 812 times out of 919 valid runs =88% so the answer is E

Problem 15

The SVM beat the regular RBF 626 times out of 791 valid runs =79% so the answer is D

Problem 16

The most common situation is E_{in} and E_{out} both go down so the answer is D

Problem 17

The most common situation is E_{in} and E_{out} going up so the answer is C

Problem 18

RBF achieves $E_{\text{in}} = 23,$ times out of 913 so the answer is A

Problem 19

Posterior:

$$\mathbb{P}(h = f|\mathcal{D})$$

$$\mathbb{P}(h = f|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|h = f)\mathbb{P}(h = f)}{\mathbb{P}(\mathcal{D})} \propto \mathbb{P}(\mathcal{D}|h = f)\mathbb{P}(h = f)$$

Since the data set is just one confirmed heart attack, $\mathbb{P}(\mathcal{D}|h = f)$, is the probability of a heart attack given we have the target function. Thus when h increases on the interval $[0,1]$ so does , $\mathbb{P}(\mathcal{D}|h = f)$ linearly,

The answer is B

Problem 20

With mean squared error we have:

$$\frac{1}{2}\frac{1}{N}\sum_{n=1}^{N}(g_1(x_n) - y_n)^2 + (g_2(x_n) - y_n)^2$$

$$E_{\text{out}_{\text{avg}}} = \frac{1}{2N}\sum_{n=1}^{N} g_1(x_n)^2 - 2g_1(x_n)y_n + y_n^2 + g_2(x_n) - 2g_2(x_n)y_n + y_n^2$$

$$= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}(g_1(x_n)^2 + g_2(x_n)^2) - (g_1(x_n)y_n + g_2(x_n)y_n) + y_n^2$$

$$E_{\text{out}}(g) = \frac{1}{N}\sum_{n=1}^{N}(g(x_n) - y_n)^2 = \frac{1}{N}\sum_{n=1}^{N} g(x_n)^2 - 2g(x_n)y_n + y_n^2$$

$$= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}\left(g_1(x_n) + g_2(x_n)\right)^2 - (g_1(x_n)y_n + g_2(x_n)y_n) + y_n^2$$

$$E_{\text{out}}(g) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}\left(\sum_{j=1}^{2} g_j(x_n)\right)^2 - \big(g_1(x_n) + g_2(x_n)\big)y_n + y_n^2$$

$$E_{\text{out}_{\text{avg}}} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}\sum_{j=1}^{2} g_j(x_n)^2 - \big(g_1(x_n) + g_2(x_n)\big)y_n + y_n^2$$

By cachy Swartz we know that $\left(\sum_{j=1}^{2} g_j(x_n)\right)^2 \leq \sum_{j=1}^{2} g_j(x_n)^2$

$$\Rightarrow E_{\text{out}_{\text{avg}}} \geq E_{\text{out}}(g)$$

So the answer C.