FINAL: 90 Minutes

Last Name: Solutions

First Name: _____

RIN: _____

Section: 4100 / 6100 (circle one)

Answer **ALL** questions.
NO COLLABORATION or electronic devices. Any violations result in an F.
NO questions allowed during the test. Interpret and do the best you can.
ALWAYS show your work and justify each answer.

GOOD LUCK!

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	TOTAL

- **You do not have time to waffle. So, don't waffle.**
- **Keep your answers precise and concise.**
- **Each question is worth 1 point.**

1. You have 30 1st graders and 30 high-school seniors. You can ask a small number of math questions to determine the best mathematician in each group. In which group is it easier to succeed?

1st graders 30 'simple' hypotheses.

2nd graders 30 "complex" hypotheses $\longrightarrow$ more variance with small dataset

$\longrightarrow$ more likely to succeed with 1st graders.

2. Van Erp got a data set from the client and produced a final hypothesis which predicted perfectly on the data. The savy client didn't know whether to be happy or not. Why?

$E_{in} \approx 0$ is only the first step.

You also need $E_{out} \approx E_{in}$ $\longleftarrow$ Client does not know this

$\therefore$ doesn't know whether to be happy or not.

3. A client has a learning problem with 10^6 data points and wants to learn a target function that is *completely unknown*. The client wants to know what you can promise her. What's your answer?

You will emerge with one of two statements:

① I failed. Here is your g and its bad $g \not\approx f$

② I succeeded and here is your $g \approx f$.

With high probability the statement you make will be true.

4. The coffee shop owner read the 10 customer feedback comments and was confused to see very polarized reviews: 5 were glowing and 5 were raging. What learning from data trap could the owner be in?

Sampling bias.

Only people with very polarized opinions tend to fill out the survey/review

The "normal" reviews are all missing

5. Give a formula for the in-sample error in logistic regression. What are the advantages of this in-sample error over some others that we could have chosen?

$$E_{in}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} \ln(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}).$$

- Based on probabilistic interpretation of $h(x)$ $\rightarrow$ max likelihood
- Convex, so easy to optimize.

6. Define a break point of a hypothesis set $\mathcal{H}$ and define the VC dimension.

k^* is a <u>break point</u> if no data set of size k^* can be shattered

<u>shatter</u>: A data set of size k^* is shattered if every dichotomy can be implemented.

<u>VC-dimension</u>: Maximum # of ~~data sets~~ data points that can be shattered

= smallest break point -1.

7. Prove that if k^* is a break point of $\mathcal{H}$, then all $k \geq k^*$ are break points.

suppose $k \geq k^*$ is not a breakpoint

$\exists$ data set of size k that can be shattered

$\rightarrow$ take k^* points from this data set

$\rightarrow$ these k^* points are shattered

$\therefore$ k^* is not a breakpoint — contradiction

$\therefore$ k is a breakpoint.

8. What is the VC-bound and why did we prove it?

$$E_{out} \leq E_{in} + \underbrace{O\left(\sqrt{\frac{1}{N}\ln m_{\mathcal{H}}(2N)}\right)}_{O\left(\sqrt{\frac{d_{vc}\ln N}{N}}\right)}$$

We proved it because it

Establishes the link theoretically between E_{in} and E_{out}

$E_{in} \approx E_{out}$ (as $N \to \infty$).

This is the second step in learning.

9. Why do we need regularization?

Regularization prevents (or helps with) overfitting to the noise.

We need it because there is <u>always</u> noise.

10. What is the difference between stochastic and deterministic noise. Which one causes overfitting?

<u>Stochastic</u>: randomness in the data. Different realizations of stochastic noise are different.

<u>deterministic</u>: arises from over complex $f \longrightarrow$ inability to model f.

<u>Both</u> cause overfitting.

11. Explain the K-RBF network. What is the functional form of the final hypothesis? What are the parameters that must be fitted to data. How do you fit the parameters?

$$g(x) = \text{sign}\left(w_0 + \sum_{i=1}^{K} w_i \, \phi\left(\frac{\|x - \mu_i\|}{r} \right) \right)$$

ϕ is the kernel.

$\mu_1 \ldots \mu_K$ are the centers (of clusters) to be learned from data.

$w_0 w_1 \ldots w_K$ are the weights to be learned from the data.

r is the data "scale"

first fit $\mu_1 \ldots \mu_K$ using Lloyds algorithm

then, given $\mu_1 \ldots \mu_K$, fit $w_0 \ldots w_K$ using any linear model algorithm, eg. PLA.

12. What is the difference between the perceptron with non-linear feature transform to K dimensions, the K-RBF network and a one-hidden-layer neural network with K hidden units? Which is most powerful?

perceptron: $g(x) = \text{sign}(w_0 + \sum_{i=1}^{K} w_i \phi_i(x))$ ϕ_i are fixed transforms.

K-RBF: $g(x) = \text{sign}\left(w_0 + \sum_{i=1}^{K} w_i \phi\left(\frac{\|x-\mu_i\|}{r}\right)\right)$ tunable transform with μ_i (unsupervised μ_i)

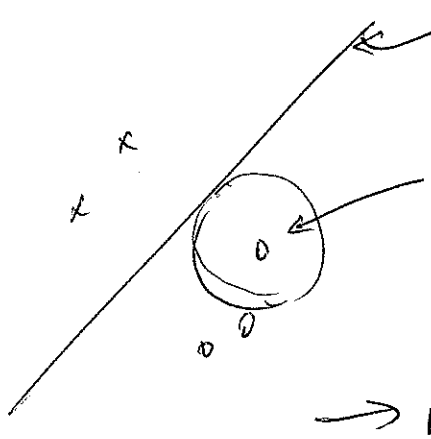
NN: $g(x) = \text{sign}(w_0 + \sum_{i=1}^{K} w_i \phi(v_i^T x))$ ← fully tunable transforms with v_i^T.

NN is most powerful.

13. Define the optimal separating hyperplane that we used in the Support Vector Machine (SVM).

Optimal seperating hyperplane separates the data while maximizing the distance to the closest data point.

14. Explain the geometric intuition for why the SVM-hyperplane is better than a random linear separator.



optimal.

maximizes the size of the error ball you can put around the data while still getting all data correct

→ maximum robustness to noise

≈ regularized.

15. *Prove* that the optimal hyperplane having maximum margin minimizes $\|\mathbf{w}\|^2$ subject to separating the data. In your proof, you should specify exactly what separating the data means.

Separate the data. $\min_n y_n(w^T x_n + b) \geq 1$

Distance of a point x_n to the hyperplane (w, b).

is given by $\dfrac{|w^T x_n + b|}{\|w\|} = d(x_n, (w,b))$.

$\therefore$ min distance is $\min_n \dfrac{|w^T x_n + b|}{\|w\|}$

$$= \frac{1}{\|w\|} \underbrace{\min_n |y_n(w^T x_n + b)|}_{=1 \text{ by separation condition}}$$

$$= \frac{1}{\|w\|}$$

$\therefore$ maximize $\dfrac{1}{\|w\|}$ s.t. $\min_n y_n(w^T x_n + b) \geq 1$.

$\Longleftrightarrow$ minimize $\|w\|^2$ s.t. $\min_n y_n(w^T x_n + b) = 1$.

this problem is mathematically equivalent to

minimize $\|w\|^2$ s.t. $y_n(w^T x_n + b) \geq 1$ for all $n = 1, \dots, N$.