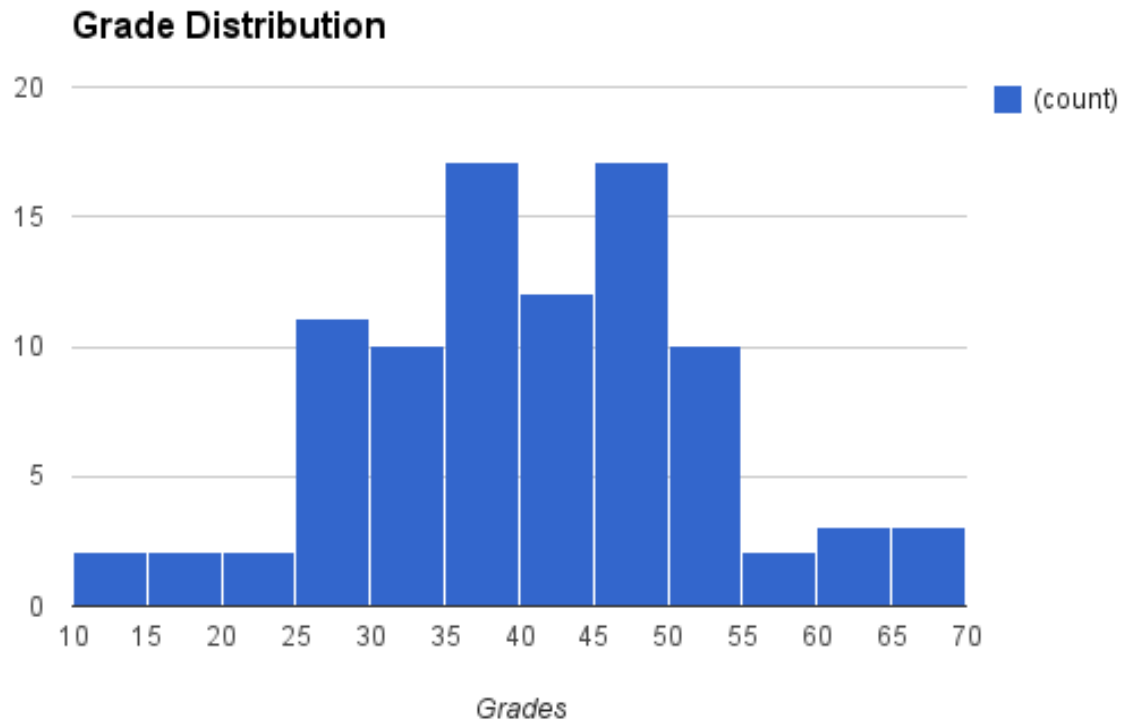
Midterm Q&A

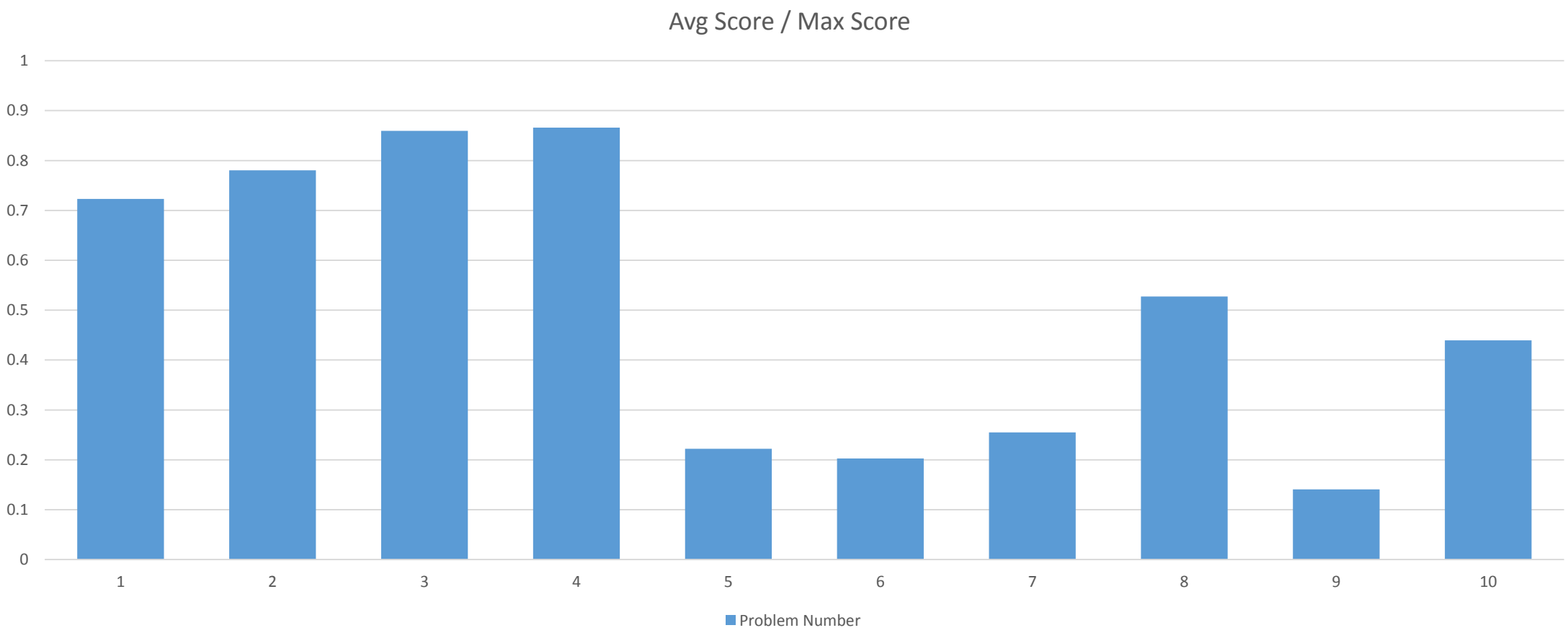
Jin Sun

Grade Distribution

- Mean: 40.28
- Stdev: 11.58
- Median: 40
- Max: 68
- Min: 10



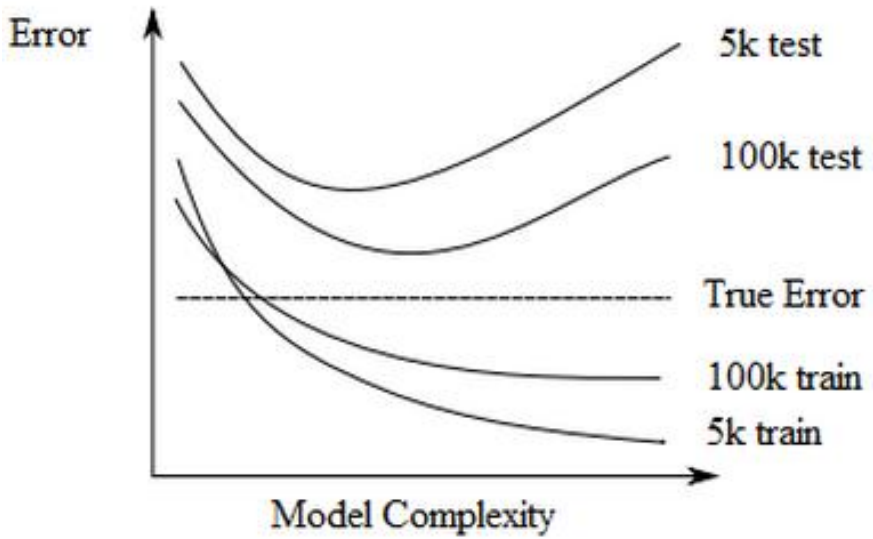
Score per Question



TAs in charge

- Q1: Jin
- Q2: Di
- Q3: Zhou
- Q4: Jin
- Q5: Jay-Yoon
- Q6: Jay-Yoon
- Q7: Manzil
- Q8: Zhou
- Q9: Manzil
- Q10: Di

Q1.1



$$f_{sample}(x) = f_{true}(x) + \epsilon$$

- The training error decreases when increasing the model complexity, while the test error decreases first but then increases due to overfitting.
- Given the same model complexity, the model has larger training samples is less likely to overfit than the one with less samples. So the two curves representing 100k samples are nearer the dash line the other two curves.

Q5

Denote by $\ell : \mathbb{R} \to RR$ a nonnegative convex differentiable function, i.e. a loss function. Moreover, denote by $\|\cdot\|_2$ the Euclidean norm. Finally, let $y_i \in \mathbb{R}$ be scalars and $x_i, w \in \mathbb{R}^d$ be vectors (i.e. $d \in \mathbb{N}$). Prove that the solution of the optimization problem

$$\underset{w}{\text{minimize}} \sum_{i=1}^{n} \ell(y_i - \langle x_i, w \rangle) + \|w\|_2$$

can be written as

$$w^* = \sum_i \alpha_i x_i \text{ for some } \alpha_i \in \mathbb{R}.$$

- Decompose w: $w = w_{//} + w_{\perp}$
  - $w_{//} \in span\{x_1, \dots, x_n\},\ w_{\perp} \in null\{x_1, \dots, x_n\}$
- $\langle x_i, w \rangle = \langle x_i, w_{//} \rangle + \langle x_i, w_{\perp} \rangle = \langle x_i, w_{//} \rangle$
- $\text{minimize}||w|| = \text{minimize}\sqrt{\|w_{//}\|^2 + \|w_{\perp}\|^2} = ||w_{//}||$
- $w^* = w_{//}$

Q6

- Primal Optimization Problem
  - $\text{minimize}_x\ f(x)$
  - Subject to $c_i(x) \leq 0$ for all i
- General Process of Obtaining Dual Form
  - Write out the Lagrange dual formulation:
    - $L(x, \alpha) = \text{maximize}_\alpha\ \text{minimize}_x\ f(x) + \alpha_i c_i(x)$
    - Subject to $\alpha \geq 0$
  - Take the derivative with respect to x, set it to zero. Solve it and plug x^* back into the dual formulation.
  - The dual formulation becomes:
    - $\text{maximize}_\alpha g(\alpha)$
    - Subject to $\alpha \geq 0$

Q7

- Log Partition function: $g(\theta) = \log \int_x \exp(\langle \phi(x), \theta \rangle) dx$

- $\nabla_\theta g(\theta) = \frac{\int_x \exp(\langle \phi(x), \theta \rangle) \cdot \phi(x) dx}{\int_x \exp(\langle \phi(x), \theta \rangle) dx} = \frac{\int_x \exp(\langle \phi(x), \theta \rangle) \cdot \phi(x) dx}{\exp(g(\theta))} =$
  $\int_x \exp(\langle \phi(x), \theta \rangle - g(\theta)) \cdot \phi(x) dx = \int_x p(x|\theta) \phi(x) dx$

Q8

- Check lecture videos
  - Sum of kernels is a kernel → Concatenating dimensions
  - Product of kernels is a kernel → Schur product theorem, Kronecker product
- Cauchy-Schwartz
  - $k(x, x')^2 = \langle \phi(x), \phi'(x) \rangle^2 \leq \|\phi(x)\|^2 \, \|\phi'(x)\|^2 = k(x, x) k(x', x')$

Q9

- Hoeffding's theorem: $P(\hat{\mu}_m - \mu > \epsilon) \leq \exp\left(-\frac{2m\epsilon^2}{c^2}\right)$

9.3 Worst Case for the Cluster

Give a lower bound on the probability that *none of the machines* in the cluster will need to perform more than a fraction of $1/M + \epsilon$ work. Hint: why can you treat the machines as if they were independent.

Solution:

$$\Pr\left(\forall m : X_m < \frac{1}{M} + \epsilon\right) = 1 - \Pr\left(\exists m : X_m > \frac{1}{M} + \epsilon\right)$$

$$\text{By union bound} \geq 1 - \sum_m \Pr\left(X_m > \frac{1}{M} + \epsilon\right) \tag{4}$$

$$\geq 1 - M \exp\left(-2N\epsilon^2\right)$$

- P(No machine overloaded) = 1 - P(At least one is overloaded)
- P(At least one is overloaded among M) ≤ M*P(single overloaded)

Q10

- Markov's Inequality: $P(X \geq \epsilon) = \frac{\mu}{\epsilon}$
- Chebyshev's Inequality: $P(|X - \mu| \geq \epsilon) = \frac{\sigma^2}{\epsilon^2}$
- Having at least 9 shoes is equivalent as having at least 10 shoes!
  - Improved bound: $P(|X - \mu| \geq \epsilon + 1)$