EE488 Special Topics in EE <Deep Learning and AlphaGo>

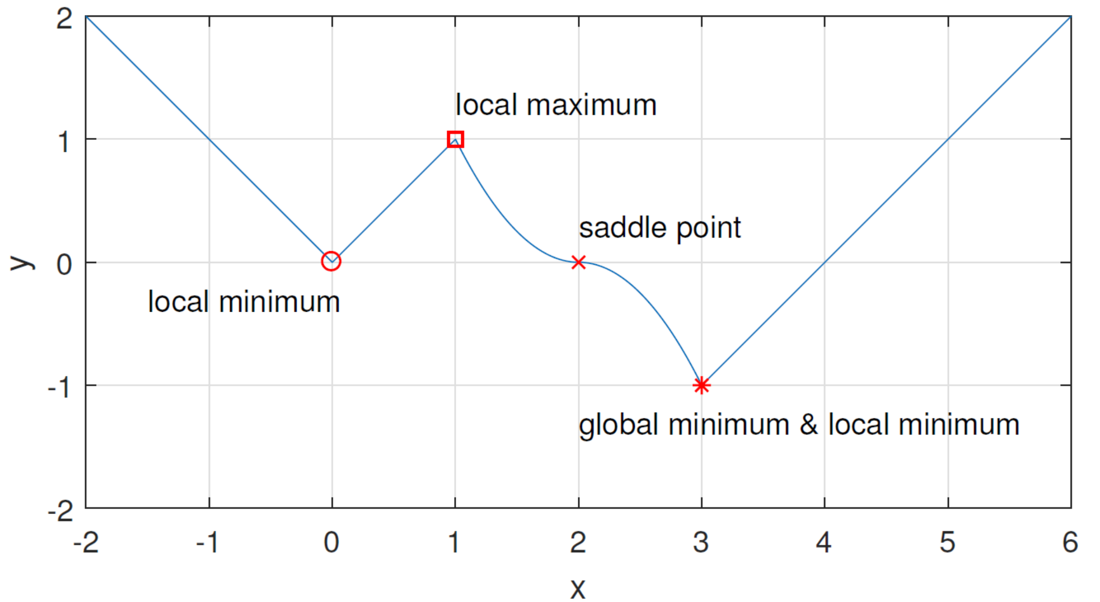
Sae-Young Chung

Some mid-term exam problems

October 23, 2017

Problem 1

1. (Local minimum, global minimum and saddle point)



Note that the global minimum is also a local minimum since the function value only increases in a small neighborhood of $x = 3$. $x = 2$ is a saddle point since $f'(2) = 0$ but it is neither local minimum or local maximum. Note that $f(x)$ is differentiable once at $x = 2$, but not twice.

Problem 6

6. (Backpropagation)

Solution)

(a)

$$\begin{aligned}
\frac{\partial J}{\partial z_i} &= -\frac{\partial}{\partial z_i} \sum_{i'} \log \sigma((2y_{i'} - 1)z_{i'}) \\
&= -\frac{\partial}{\partial z_i} \log \sigma((2y_i - 1)z_i) \\
&= (1 - 2y_i)(1 - \sigma((2y_i - 1)z_i)) \\
&= \begin{cases} \hat{y}_i & \text{if } y_i = 0 \\ \hat{y}_i - 1 & \text{if } y_i = 1 \end{cases} \\
&= \hat{y}_i - y_i,
\end{aligned}$$

where the third equality follows since $\frac{\partial}{\partial z} \log \sigma(\alpha z) = \frac{\partial}{\partial z} \log \frac{1}{1+\exp(-\alpha z)} = \frac{\alpha \exp(-\alpha z)}{1+\exp(-\alpha z)} = \alpha(1 - \sigma(\alpha z))$ and the fourth equality follows by considering two cases $y_i = 0$ or 1 and by using $1 - \sigma(-z) = \sigma(z)$.

(b) Using the result in (a), we get

$$\nabla_{\mathbf{z}} J = \hat{\mathbf{y}} - \mathbf{y}.$$

Recap – Example 1 (Lecture Notes #6)

- Training set: $(\mathbf{x}, \mathbf{y})$
  - $\mathbf{x}$: $m \times 1$ vector of m inputs for training
  - $\mathbf{y}$: $m \times 1$ vector of m outputs for training
- Assume l layers with 1 neuron in each layer
  - First layer output: $\mathbf{h}_1 = \phi^{(1)}(\mathbf{x} w_1)$ (assume no bias for simplicity) (define also $\mathbf{h}_0 = \mathbf{x}$)
    - w_1: weight of the first layer
    - $\phi^{(1)}(\cdot)$: activation function of the first layer
    - $\mathbf{h}_1$: $m \times 1$ vector containing m outputs of the first layer
  - Second layer output: $\mathbf{h}_2 = \phi^{(2)}(\mathbf{h}_1 w_2)$
    - w_2: weight of the second layer
    - $\phi^{(2)}(\cdot)$: activation function of the second layer
    - $\mathbf{h}_2$: $m \times 1$ vector containing m outputs of the second layer
  - l-th layer output: $\mathbf{h}_l = \phi^{(l)}(\mathbf{h}_{l-1} w_l)$ (output layer)
- Cost

$$J(w_1, w_2, \ldots, w_l) = \frac{1}{2m} \|\mathbf{y} - \mathbf{h}_l\|^2$$

Recap – Example 1 (Lecture Notes #6)

- Goal: to calculate gradients $\frac{\partial J}{\partial w_1}, \frac{\partial J}{\partial w_2}, \ldots, \frac{\partial J}{\partial w_l}$
- Let's define $\mathbf{u}_k = \mathbf{h}_{k-1} w_k$ and $\mathbf{g}_k = \nabla_{\mathbf{u}_k} J$, $k = 1, \ldots, l$, then $\mathbf{g}_l = \frac{1}{m}\phi^{(l)'}(\mathbf{u}_l) \odot (\mathbf{h}_l - \mathbf{y})$ and $\mathbf{g}_{k-1} = \phi^{(k-1)'}(\mathbf{u}_{k-1}) \odot \mathbf{g}_k w_k$, $k = 2, \ldots, l$ since

$$g_{l,i} = \frac{\partial J}{\partial u_{l,i}} = \frac{\partial}{\partial u_{l,i}} \frac{1}{2m} \|\mathbf{y} - \phi^{(l)}(\mathbf{u}_l)\|^2 = \frac{\partial}{\partial u_{l,i}} \frac{1}{2m}(y_i - \phi^{(l)}(u_{l,i}))^2 = \frac{1}{m}\phi^{(l)'}(u_{l,i})(\phi^{(l)}(u_{l,i}) - y_i)$$

$$g_{l-1,i} = \frac{\partial J}{\partial u_{l-1,i}} = \sum_{i'} \frac{\partial u_{l,i'}}{\partial u_{l-1,i}} \frac{\partial J}{\partial u_{l,i'}} = \phi^{(l-1)'}(u_{l-1,i}) w_l g_{l,i}$$

$$\vdots$$

$$g_{1,i} = \phi^{(1)'}(u_{1,i}) w_2 g_{2,i}$$

- $\frac{\partial J}{\partial w_k} = \mathbf{h}_{k-1}^T \mathbf{g}_k$, $k = 1, \ldots, l$ since

$$\frac{\partial J}{\partial w_k} = \sum_{i'} \frac{\partial u_{k,i'}}{\partial w_k} \frac{\partial J}{\partial u_{k,i'}} = \sum_{i'} h_{k-1,i'} g_{k,i'} = \mathbf{h}_{k-1}^T \mathbf{g}_k$$

6. (Backpropagation – 10 points) Let's consider a two-layer neural network given by

$$h = \phi(w_1 x + b_1)$$
$$\hat{y} = \sigma(w_2 h + b_2),$$

where x is the input, h is the output of the first layer, $\hat{y}$ is the final output, $\phi(\cdot)$ is an activation function for the hidden layer and $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the sigmoid function. x, w_1, b_1, h, w_2, b_2 and $\hat{y}$ are all scalars. Let $(\mathbf{x}, \mathbf{y})$ denote the set of m training examples, where $\mathbf{x}$ is the $m \times 1$ vector of m inputs for training and $\mathbf{y}$ is the $m \times 1$ vector of m outputs for training. Let x_i and y_i denote the i-th element of $\mathbf{x}$ and $\mathbf{y}$, respectively. Let's define $\mathbf{u} = w_1\mathbf{x} + b_1$, $\mathbf{h} = \phi(\mathbf{u})$, $\mathbf{z} = w_2\mathbf{h} + b_2$, and $\hat{\mathbf{y}} = \sigma(\mathbf{z})$. Let u_i, h_i, z_i, and $\hat{y}_i$ denote the i-th element of $\mathbf{u}$, $\mathbf{h}$, $\mathbf{z}$, and $\hat{\mathbf{y}}$, respectively. Assume the elements of $\mathbf{y}$ are binary, i.e., 0 or 1. Assume the cost function J is defined using the cross entropy, i.e., $J = -\sum_{i=1}^{m} \log \sigma((2y_i - 1)z_i)$. Let $\mathbf{f}$ and $\mathbf{g}$ denote $\nabla_{\mathbf{u}} J$ and $\nabla_{\mathbf{z}} J$, respectively and let f_i and g_i denote the i-th element of $\mathbf{f}$ and $\mathbf{g}$, respectively.

(b) What is $\nabla_{\mathbf{z}} J$? Write your answer below. Your answer should be expressed in terms of vectors we defined so far, e.g., $\mathbf{x}$, $\mathbf{h}$, $\mathbf{y}$, etc. Make your final expression as simple as possible. You will get 0 point if the final expression is not of the simplest form.

(e) Derive $\frac{\partial J}{\partial w_2}$, $\frac{\partial J}{\partial b_2}$, $\frac{\partial J}{\partial w_1}$, and $\frac{\partial J}{\partial b_1}$ and write your answers in the boxes below. Make your final expression as simple as possible. For each item, you will get 0 point if the final expression is not of the simplest form or if you do not provide detailed derivation steps. Your final expression should only contain vectors such as $\mathbf{x}$, $\mathbf{h}$, $\mathbf{y}$, etc.

Problem 8

8. (Overfitting or underfitting)

(a) *An estimator* $\hat{\theta}(x) = \frac{0.95+0.1x}{2}$ *that estimates* $0 \le \theta \le 1$ *from* x *following Bernoulli(θ)*

Solution) This is an example of underfitting since $\hat{\theta}(x)$ is either 0.475 (when $x = 0$) or 0.525 (when $x = 1$), which is close to a trivial estimator that always outputs 0.5. Such a trivial estimator and the estimator in this problem would be in the underfitting regime whereas $\hat{\theta}(x) = x$ would be in the overfitting regime. $\hat{\theta}(x) = \frac{1+2x}{4}$ would achieve a good balance between underfitting and overfitting.

(b) *Someone thinks Earth is sitting on four giant elephants.*

Solution)

Overfitting. Why four? Why elephants? There are infinitely many other possibilities and you are simply believing what you chose to believe based on your limited experience.

Recap – Estimation

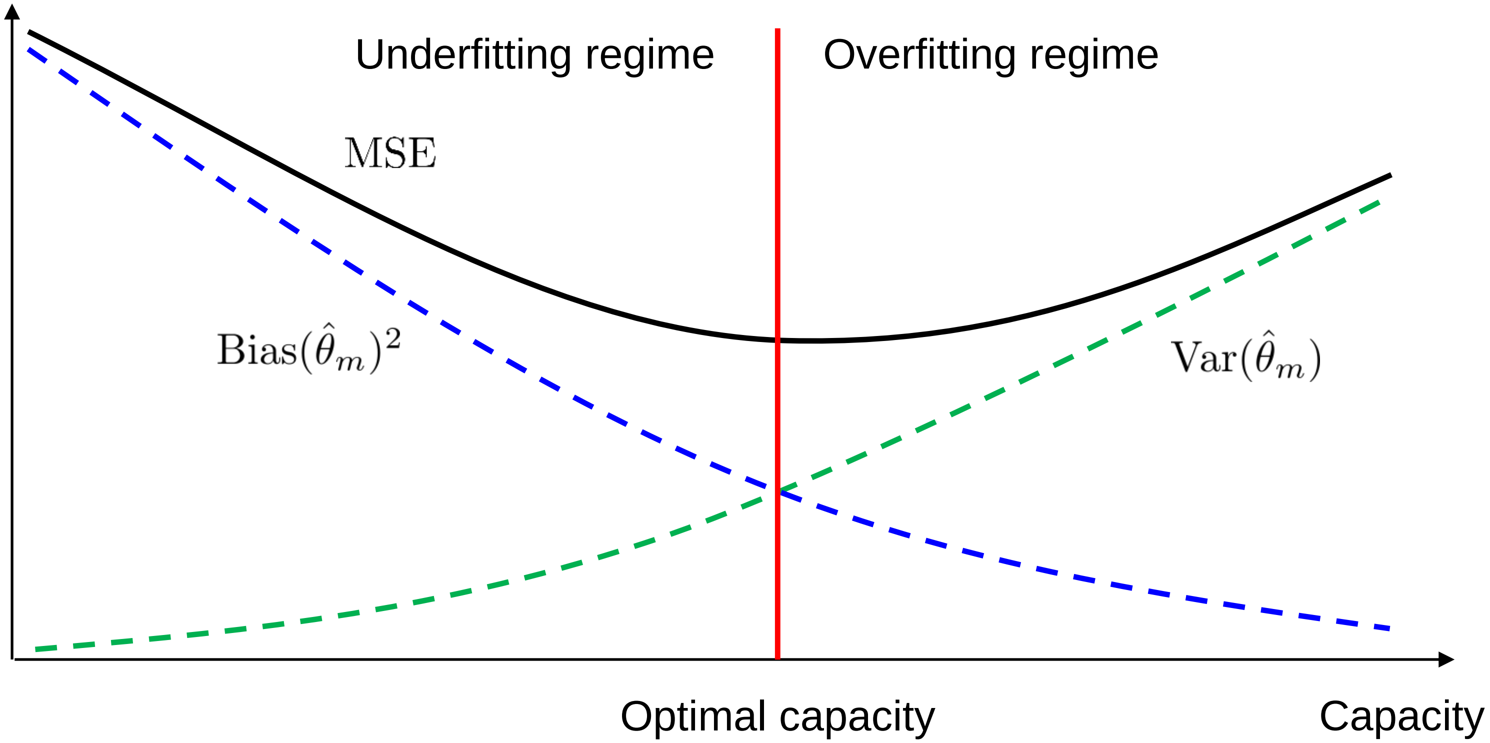
- Point estimation: estimation of a quantity of interest, say θ
  - $\{x^{(1)}, \ldots, x^{(m)}\}$: m i.i.d. data points generated by p_θ
  - $\hat{\theta}_m = g(x^{(1)}, \ldots, x^{(m)})$
  - Cf) function estimation: estimation of a function $f(x)$ based on samples of (x, y)
- Bias of an estimator: $\mathrm{bias}(\hat{\theta}_m) = \mathbb{E}(\hat{\theta}_m) - \theta$
- Variance of an estimator: $\mathrm{Var}(\hat{\theta}_m)$
- Mean squared error (MSE)

$$\begin{aligned}
\mathrm{MSE} &= \mathbb{E}[(\hat{\theta}_m - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta}_m - \mathbb{E}(\hat{\theta}_m) + \mathbb{E}(\hat{\theta}_m) - \theta)^2] \\
&= \mathrm{Var}(\hat{\theta}_m) + 2\mathbb{E}[\hat{\theta}_m - \mathbb{E}(\hat{\theta}_m)]\mathbb{E}[\mathbb{E}(\hat{\theta}_m) - \theta] + (\mathbb{E}(\hat{\theta}_m) - \theta)^2 \\
&= \mathrm{Var}(\hat{\theta}_m) + 2\{\mathbb{E}(\hat{\theta}_m) - \mathbb{E}(\hat{\theta}_m)\} \cdot \{\mathbb{E}(\hat{\theta}_m) - \theta\} + \mathrm{Bias}(\hat{\theta}_m)^2 \\
&= \mathrm{Var}(\hat{\theta}_m) + \mathrm{Bias}(\hat{\theta}_m)^2
\end{aligned}$$

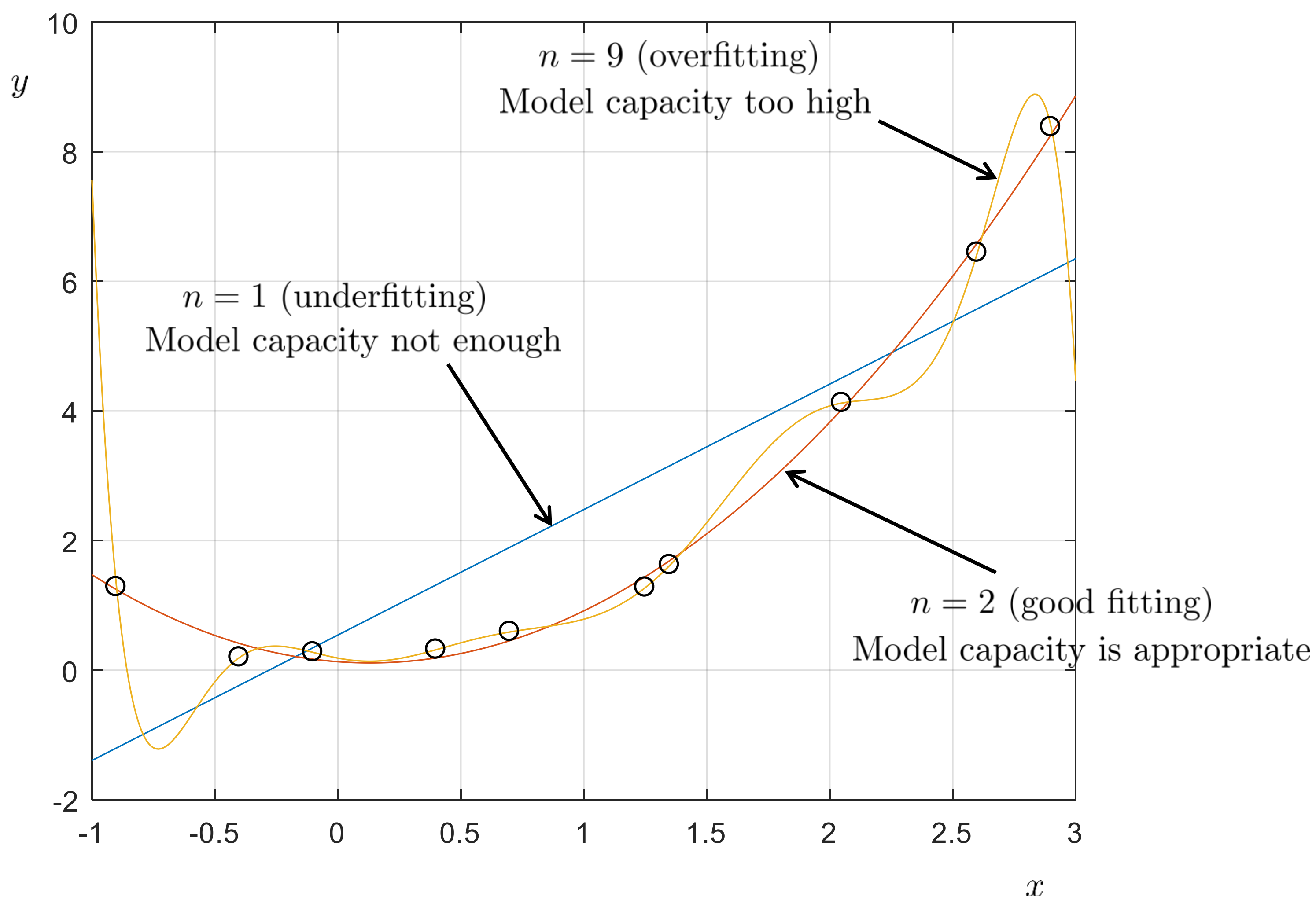
Problem 8 (a)

- $\hat{\theta}(x) = x$
  - Bias: $\mathbb{E}[\hat{\theta}(x)] - \theta = \mathbb{E}[x] - \theta = \theta - \theta = 0$
  - Variance: $\mathbb{E}[(\hat{\theta}(x) - \mathbb{E}[\hat{\theta}(x)])^2] = \mathbb{E}[(x-\theta)^2] = (0-\theta)^2(1-\theta) + (1-\theta)^2\theta = \theta(1-\theta)$
- $\hat{\theta}(x) = \frac{1}{2}$
  - Bias: $\mathbb{E}[\hat{\theta}(x)] - \theta = \frac{1}{2} - \theta$
  - Variance: $\mathbb{E}[(\hat{\theta}(x) - \mathbb{E}[\hat{\theta}(x)])^2] = 0$
- $\hat{\theta}(x) = \frac{1+2x}{4} = \begin{cases} \frac{1}{4} & \text{if } x = 0 \\ \frac{3}{4} & \text{if } x = 1 \end{cases}$
  - Bias: $\mathbb{E}[\hat{\theta}(x)] - \theta = \frac{1+2\theta}{4} - \theta = \frac{1-2\theta}{4}$
  - Variance: $\mathbb{E}[(\hat{\theta}(x) - \mathbb{E}[\hat{\theta}(x)])^2] = \frac{\theta(1-\theta)}{4}$ (see solution set #2 of last year)
  - Achieves a good balance between the above two since $0 \le |\frac{1-2\theta}{4}| \le |\frac{1}{2} - \theta|$ and $0 \le \frac{\theta(1-\theta)}{4} \le \theta(1-\theta)$

Recap

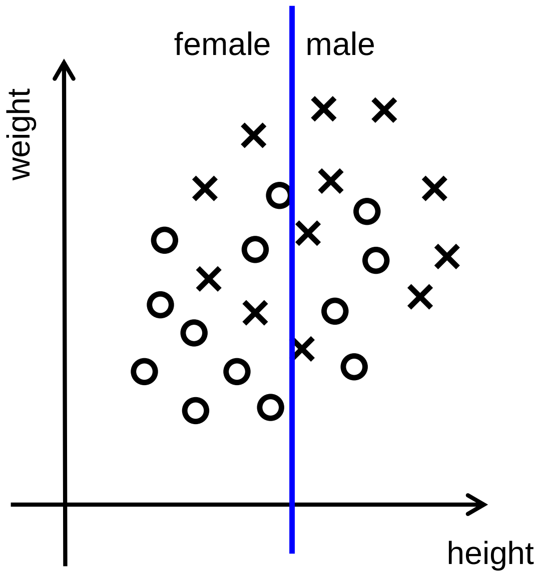

Underfitting regime
Overfitting regime
MSE
$\mathrm{Bias}(\hat{\theta}_m)^2$
$\mathrm{Var}(\hat{\theta}_m)$
Optimal capacity
Capacity

Recap – Underfitting & Overfitting

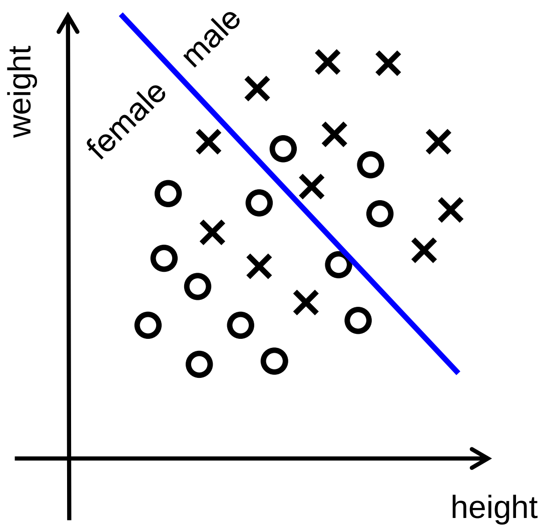


Recap – Overfitting & Underfitting in Classification

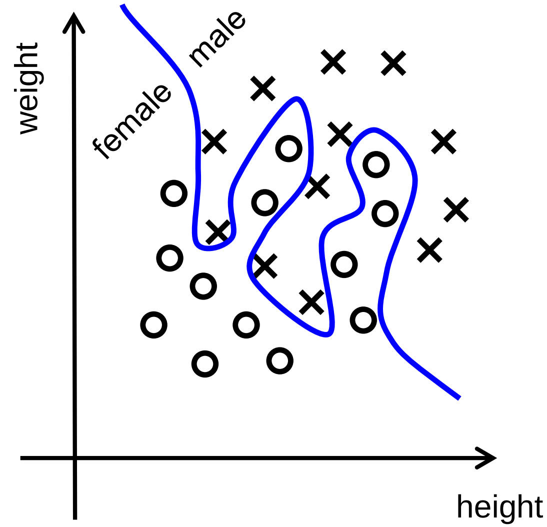
Underfit

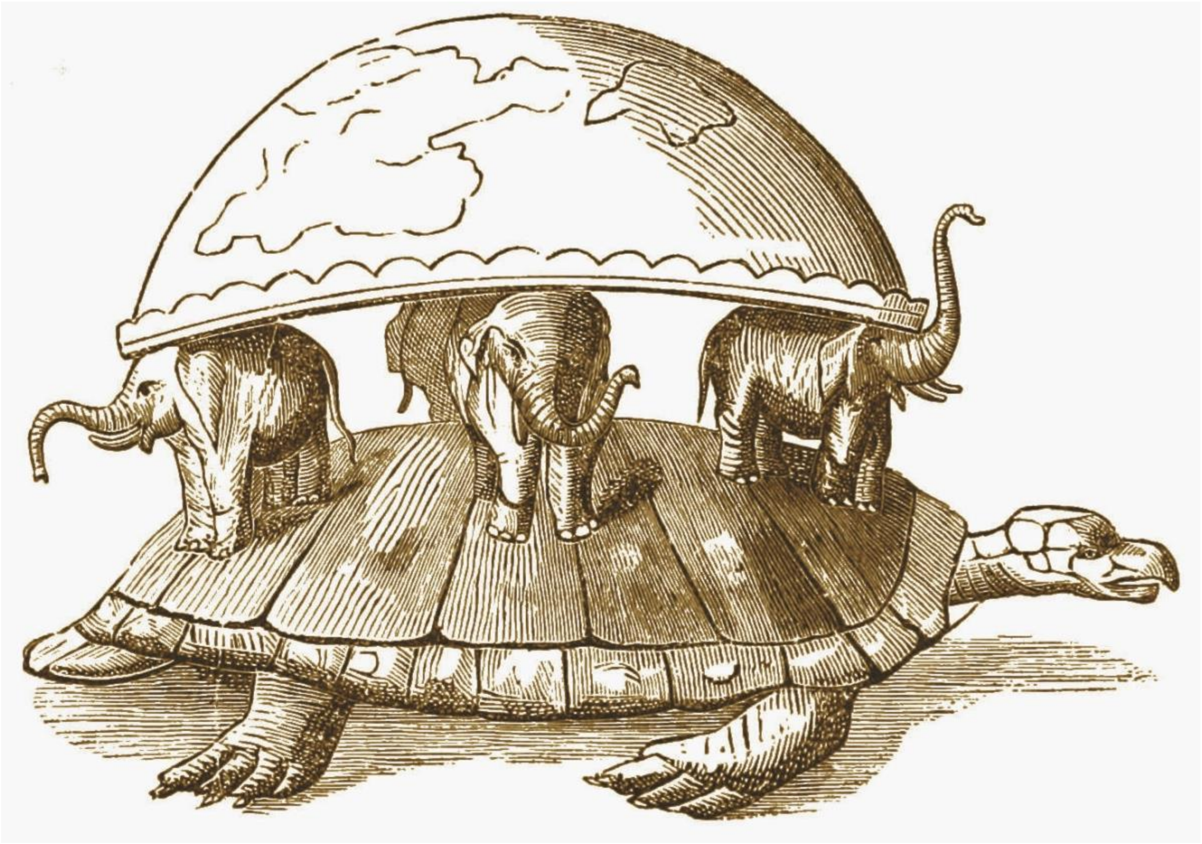


Good fit



Overfit



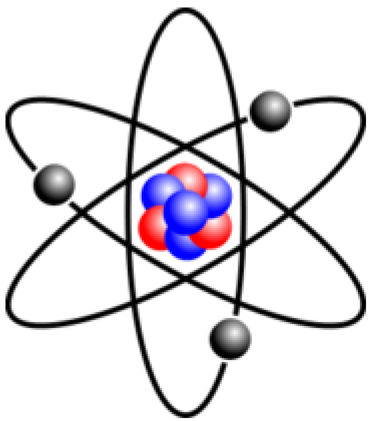


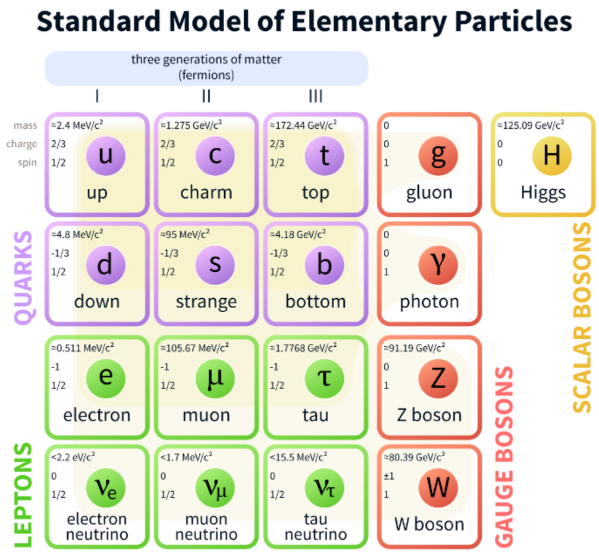
Answer

Underfitting or Overfitting?

Four Elements, Thomas Vogel

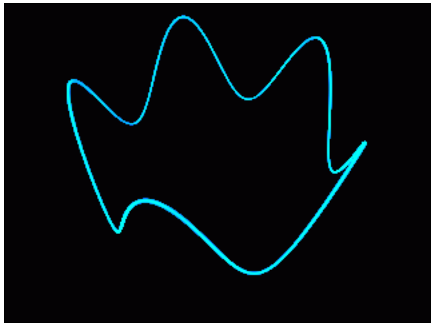
Answer (Not Found Yet)

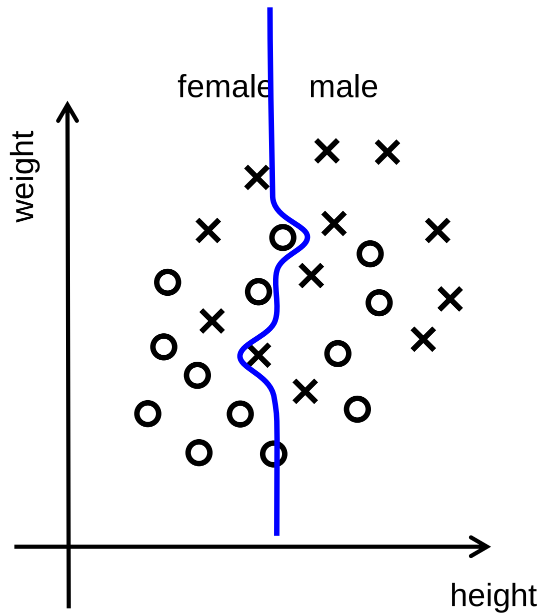


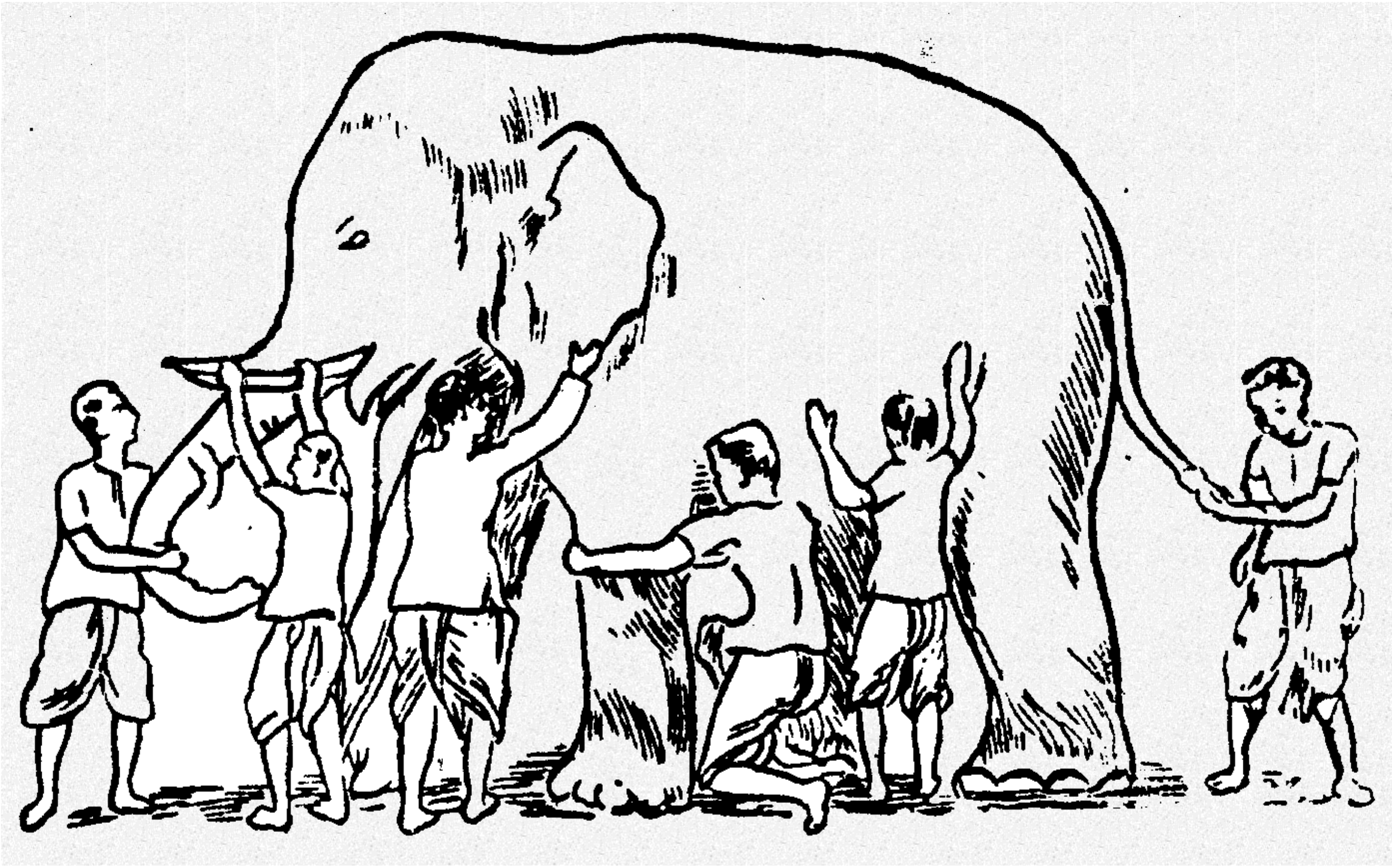


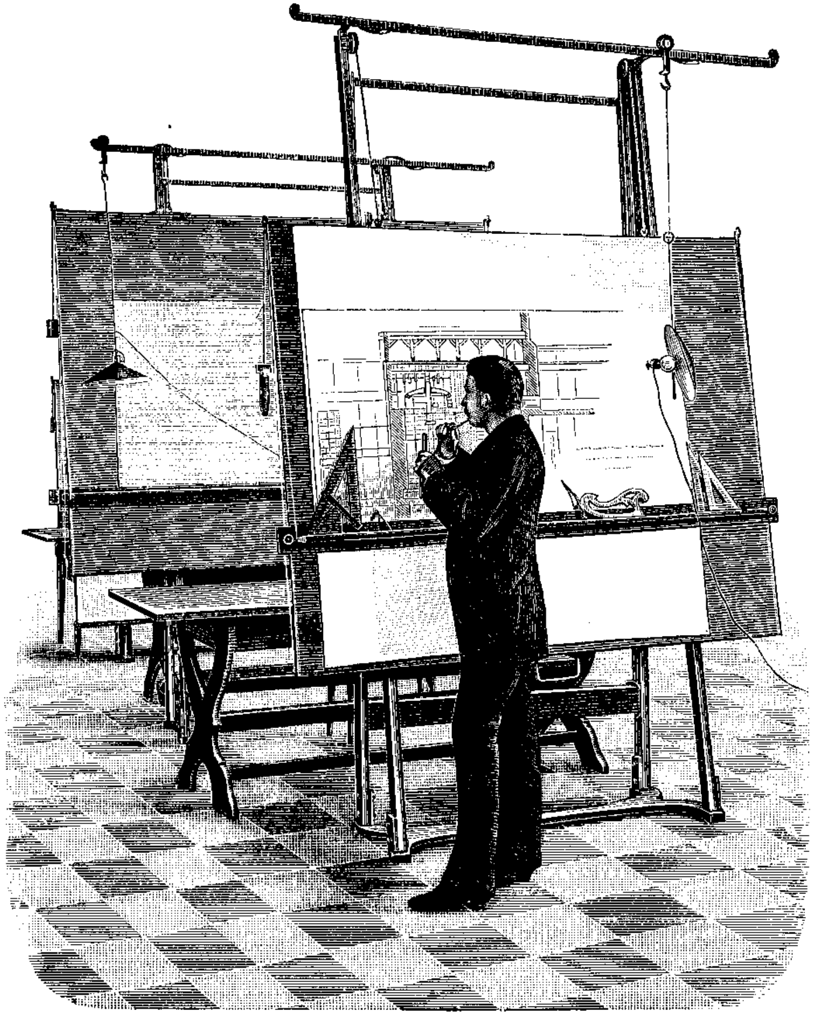
Source: Wikipedia

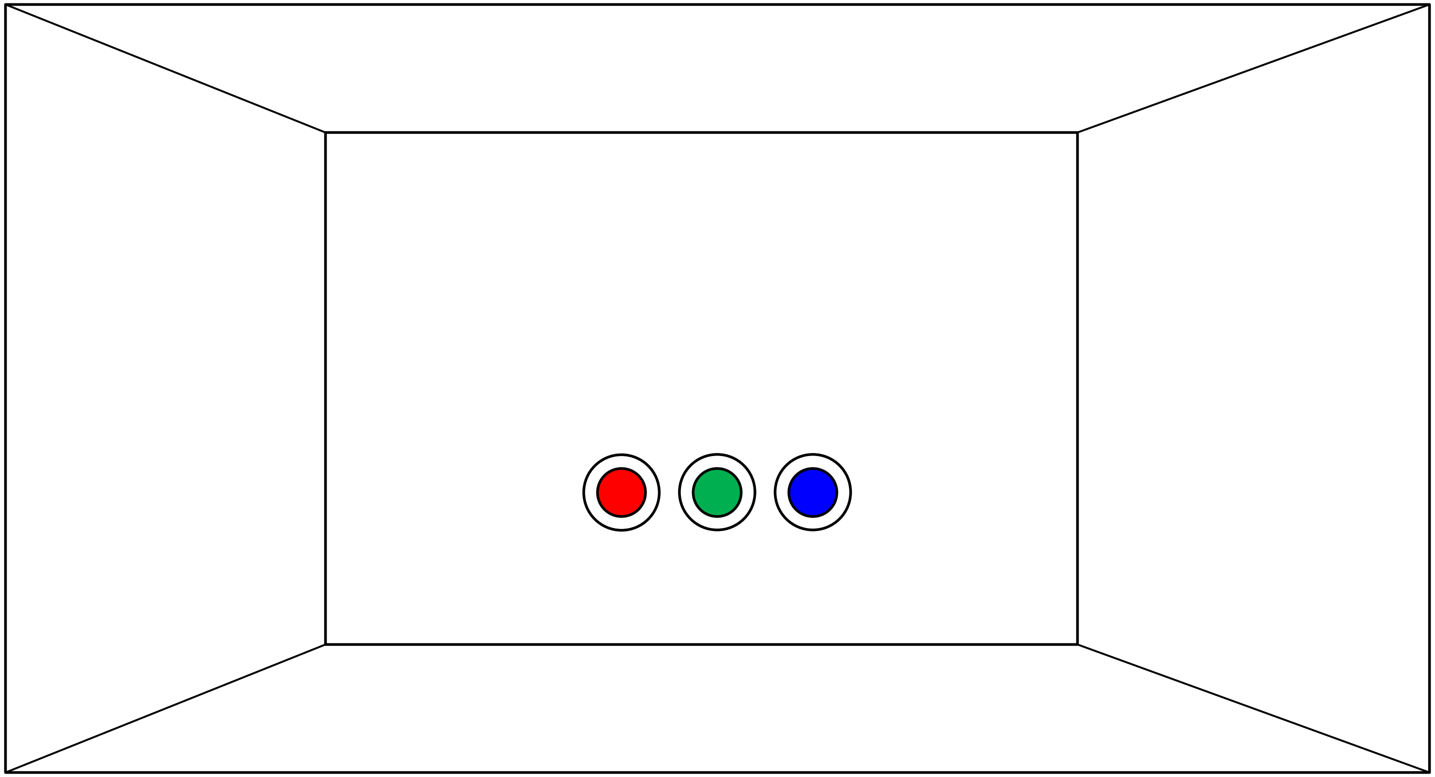
?




female
male
weight
height



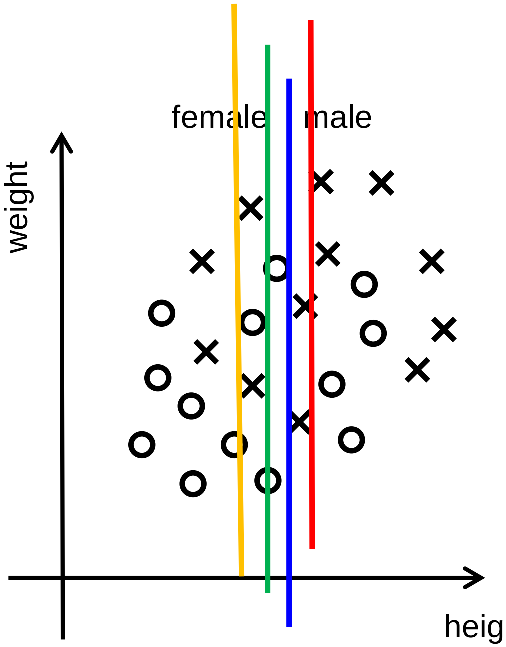




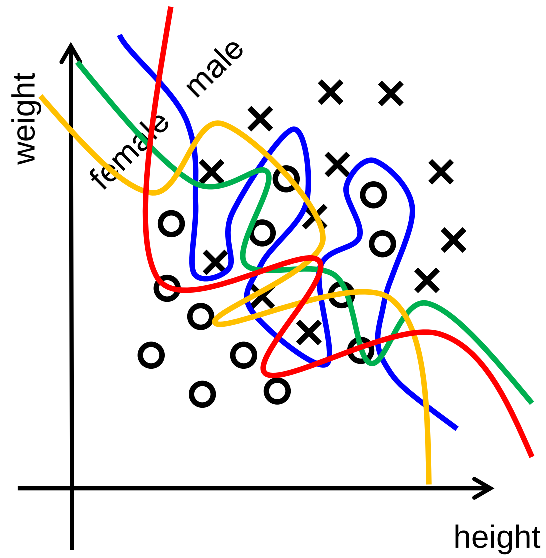
The Subtle Art of Not Giving a F*ck: A Counterintuitive Approach to Living a Good Life by Mark Manson, pp. 120-122

Which One is Better?

Multiple underfitted agents



Multiple overfitted agents



Photograph by Yukihiro Fukuda

Mid-term Stat

- Mean: 26.8
- Median: 27
- Stdev: 6.5
- Best score: 39 out of 42 (1 student)
- Distribution (128 students)