CSDS503 / COMP552 – Advanced Machine Learning

Faizad Ullah

The Gold Labels

The Output of a Classifier

#	Height (inches)	Weight (kgs)	B.P. Sys	B.P. Dia	Heart disease	
	$\vec{x}$				y	$h(\vec{x})$
1	62	70	120	80	No	No
2	72	90	110	70	No	Yes
3	74	80	130	70	No	No
4	65	120	150	90	Yes	Yes
5	67	100	140	85	Yes	No
6	64	110	130	90	No	Yes
7	69	150	170	100	Yes	Yes
8	75	127	160	95	Yes	No
9	66	66	135	90	Yes	Yes

- $\boldsymbol{y}$: Gold standards / Gold labels / Ground truth
- $\boldsymbol{h(\vec{x})}$: Predicted labels

(How) can we trust the y?

- We need to assess the agreement between two raters.
    1. Cohen's Kappa (κ) is a statistical measure used to quantify the level of agreement between two raters
    2. Fleiss' kappa can be used to assess the inter-rater agreement between two or more raters
    3. Krippendorff's Alpha is a measure for assessing inter-rater reliability

(How) can we trust the y?

- Objective data gathered from the real world:
  - Measurement of heights, weights, weather phenomena
    - Aberrations, sensor malfunctions and errors
- Subjective data, annotated by humans (experts or community)
  - Tagging emotions, SPAM/HAM, sentiment, misinformation
    - Errors of judgement, biases, human error
- Use multiple sources for each label
  - Multiple sensors measuring the same phenomenon
  - Multiple humans annotating the same data
    - Inter-annotator agreements
  - Resource intense
    - Partial overlap of data
- Is annotator agreement the answer to all our worries?
  - Distribution of annotators
  - Chance agreements

Intuition of Agreement (and Randomness)

- **Annotator1 (A1)** and **Annotator2 (A2)** are deciding if posts should be deleted from a social media platform.
- Most of the posts should not be deleted.
- A1 is not an expert and just randomly selects 1% of posts.
- A2 is an expert and very carefully selects 1% of posts.
- A1 and A2 still agree on at least 98% of posts (that need not be deleted).
- So, what is the problem?
- The problem is that A1 was just picking posts at random. Because most posts are not deleted, A1 and A2 agree most of the time.
- But that does not mean that A1 and A2 are using similar ratings.
- Cohen's Kappa corrects for this – It would be approximately zero in this case.

Gold Labels, Annotators and Agreement

Utterances	Ann1	Ann2	Raw Agreement	Ann Rand	Agreement (A1, Rand)	Agreement (A2, Rand)
S1	+	+	1	-	0	0
S2	-	-	1	+	0	0
S3	+	-	0	+	1	0
S4	-	+	0	-	1	0
…	…	…	…	…	…	…
			Sum of Agreements		Sum of chance agreements	Sum of chance agreements

- Raw (observed) agreements
- Chance agreements
- *Cohen's Kappa* ($\kappa = \frac{P_o - P_e}{1 - P_e} = 1 - \frac{1 - P_o}{1 - P_e}$)

$$\kappa = \frac{P_o - P_e}{1 - P_e} = 1 - \frac{1 - P_o}{1 - P_e}$$

- P_o is the probability of actual agreement
- P_e is the expected probability of "agreement by chance"- the probability of the agreement if the rating is random
- The numerator is the difference between the observed agreement and chance agreement
- The denominator is the difference between the maximum possible observed agreement (i.e., 1: 100% agreement) and the chance agreement
- As the denominator is just a normalizer, observe that κ is high when the difference between the observed and chance agreement is the highest
  - For a good model, the observed difference and the maximum difference are close to each other, and Cohen's kappa is close to 1.
  - For a random model, the overall accuracy is all due to random chance, the numerator is 0, and Cohen's kappa is 0.
  - Cohen's kappa could also theoretically be negative. Then, the overall accuracy of the model would be even lower than what could have been obtained by a random guess.

Cohen’s Kappa

Suppose that you were analyzing data related to a group of 50 people applying for a grant. Each grant proposal was read by two readers and each reader either said "Yes" or "No" to the proposal. Suppose the disagreement count data were as follows, where A and B are readers.

		B	
		Yes	No
A	Yes	a	b
	No	c	d

- The observed proportionate agreement is:

$$p_o = \frac{a+d}{a+b+c+d} = \frac{20+15}{50} = 0.7$$

- To calculate p_e (the probability of random agreement) we note that:
  - Reader A said "Yes" to 25 applicants and "No" to 25 applicants. Thus reader A said "Yes" 50% of the time.
  - Reader B said "Yes" to 30 applicants and "No" to 20 applicants. Thus reader B said "Yes" 60% of the time.
- So the expected probability that both would say yes at random is:

$$p_{\text{Yes}} = \frac{a+b}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d} = 0.5 \times 0.6 = 0.3$$

		B	
		Yes	No
A	Yes	20	5
	No	10	15

- Similarly,

$$p_{\text{No}} = \frac{c+d}{a+b+c+d} \cdot \frac{b+d}{a+b+c+d} = 0.5 \times 0.4 = 0.2$$

- Overall random agreement probability is the probability that they agreed on either Yes or No, i.e.:

$$p_e = p_{\text{Yes}} + p_{\text{No}} = 0.3 + 0.2 = 0.5$$

- So now applying our formula for Cohen's Kappa we get:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

Kappa value interpretation Landis & Koch (1977):

<0 No agreement

0 — .20 Slight

.21 — .40 Fair

.41 — .60 Moderate

.61 — .80 Substantial

.81–1.0 Perfect

Cohen’s Kappa

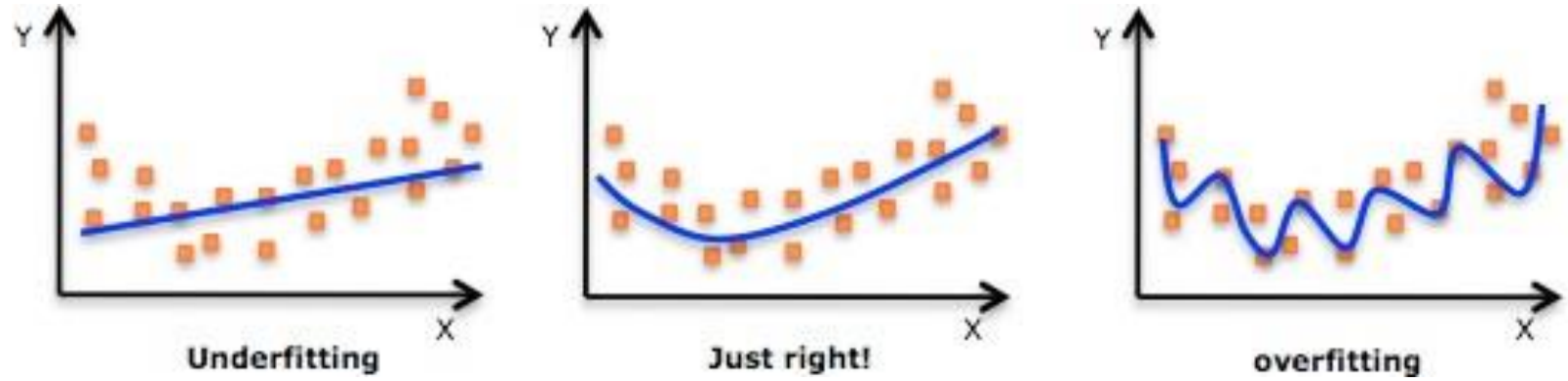
- Reliability: Are both measurements giving similar results?

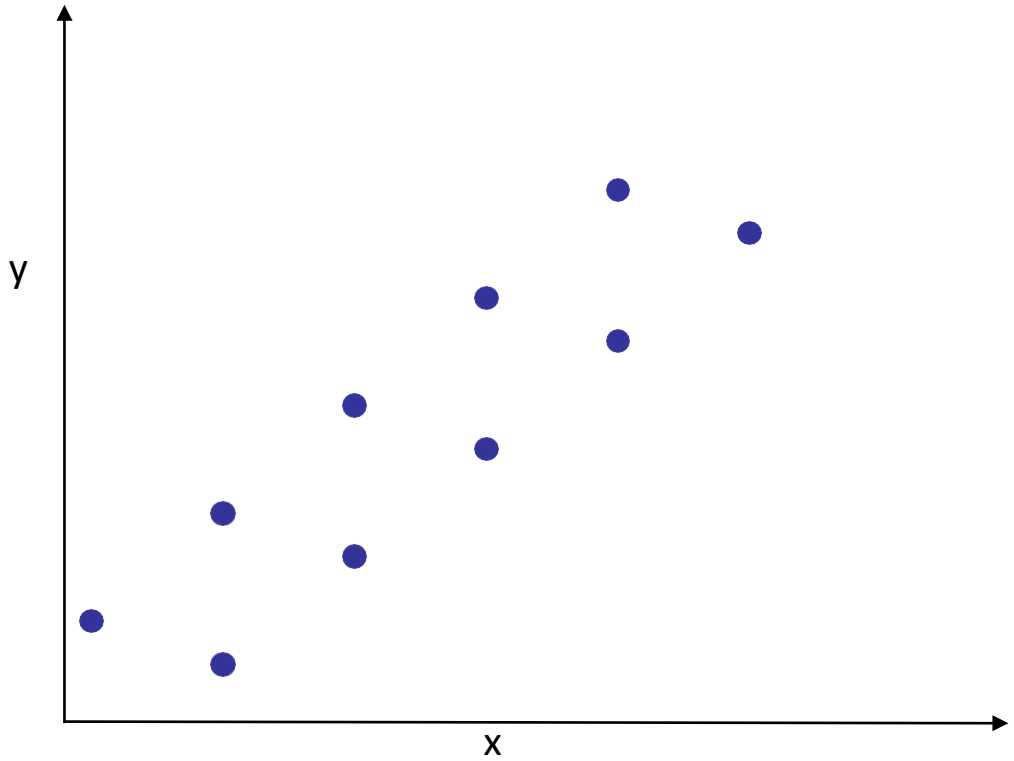
- Validity: Are the raters measuring the right thing?

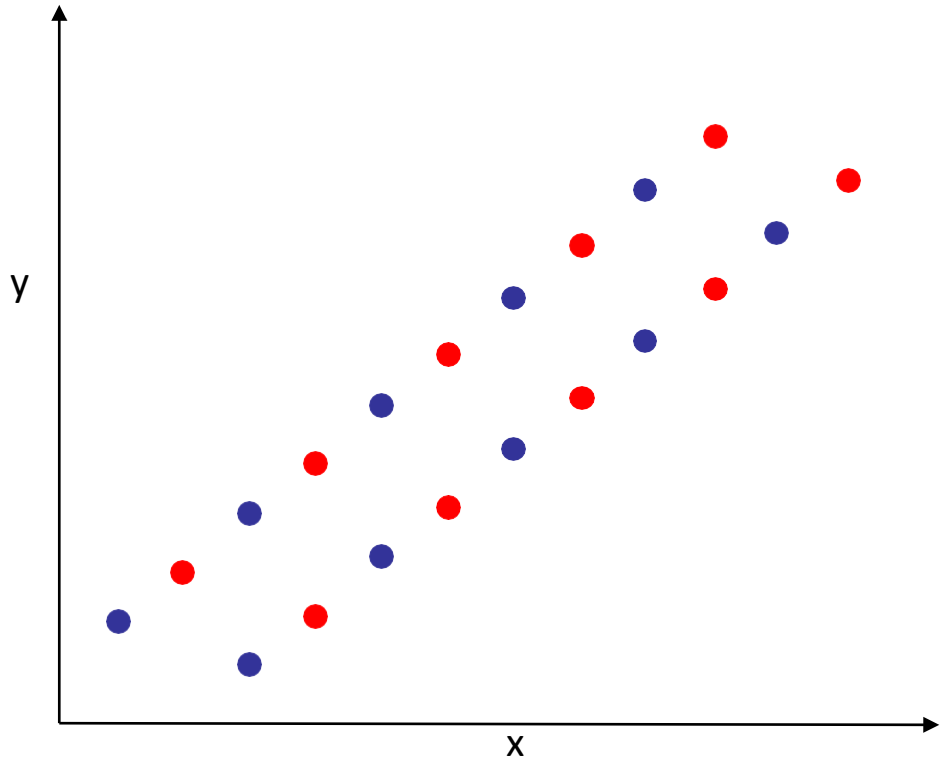
1. "I feel so Normal.“
2. "Finally got my coffee fix! Ready to conquer the day.“
3. "Another night, another sleepless struggle“
4. "Feeling like I’m just a fun to everyone around me.“
5. "Life may be hard sometimes, but I’m staying positive.“
6. "I’m so tired of pretending everything is okay.“
7. "It’s been really tough getting through each day lately.“
8. "Sometimes, I wonder if things will ever get better.“
9. "Starting the week with a fresh mindset and new goals.“
10. "Even when life is tough, I’m grateful for the journey.“
11. "Reflecting on the positives in life."
12. "Nothing feels real anymore. It’s like I’m just floating through life.“
13. "The weight of everything is crushing me.“
14. "It feels like I’m stuck in a hole I can’t get out of.“
15. "Can’t wait for the weekend, got some exciting plans!“

16. "I hope tomorrow feels a little easier."
17. "Looking forward to a weekend of relaxation and fun!"
18. "Excited to start my new project at work tomorrow!"
19. "Can't seem to shake this feeling of sadness, no matter what I do."
20. "Grateful for the little things in life."
21. "Taking it one day at a time, focusing on the good."
22. "Sunshine and coffee make everything better!"
23. "Finally feeling like myself again after a tough week."
24. "Had a wonderful time with friends last night."
25. "Today was a good day. Feeling optimistic about the future."
26. "Trying my best to keep going, but it's so hard."
27. "Just want to be alone with my thoughts for a while."
28. "I don't even remember what it feels like to be happy."
29. "Sometimes, it feels like no one would notice if I just vanished."

Overfitting – Bias and Variance

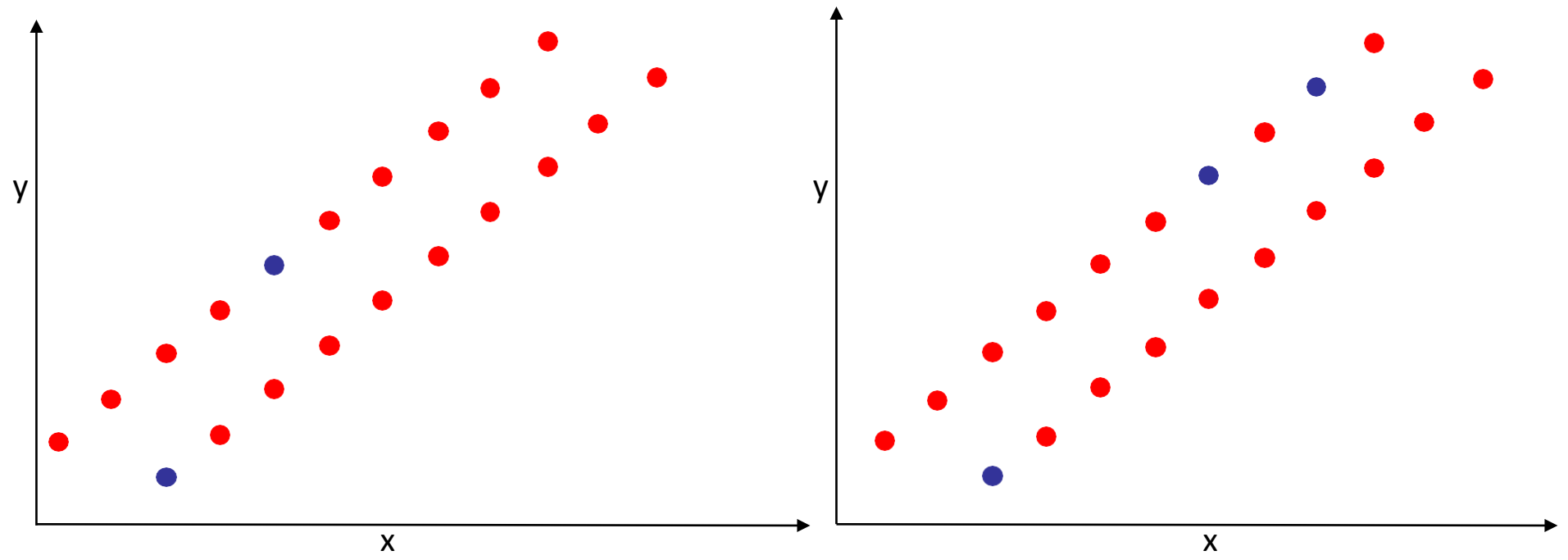

Y
X
Underfitting
Y
X
Just right!
Y
X
overfitting


y
x


y
x

http://www.petanimalexpo.co.nz/

The fitting problem



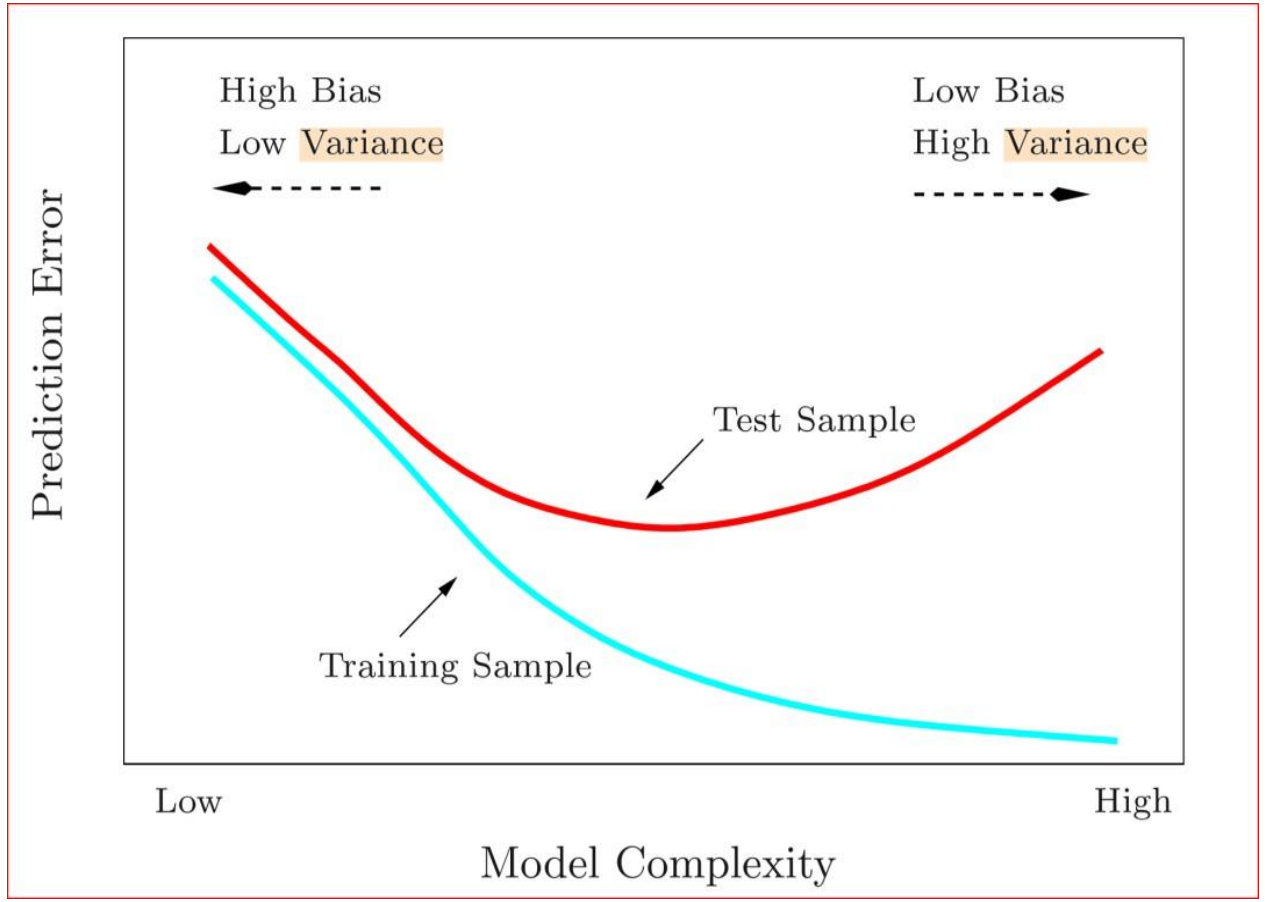
Bias and Variance

- **Bias** is the difference between the average prediction of our model and the correct value which we are trying to predict.
  - If the average predicted values are far off from the actual values, then the bias is high.
  - Model with high bias pays **little attention to the training data** and oversimplifies (presumes a lot about) the model.
  - High bias causes algorithm to **miss relevant relationship between input and output variable**.
  - When a model has a high bias then it implies that the model is too simple and does not capture the complexity of data thus **underfitting** the data.
  - It leads to high error on **training and test data**.

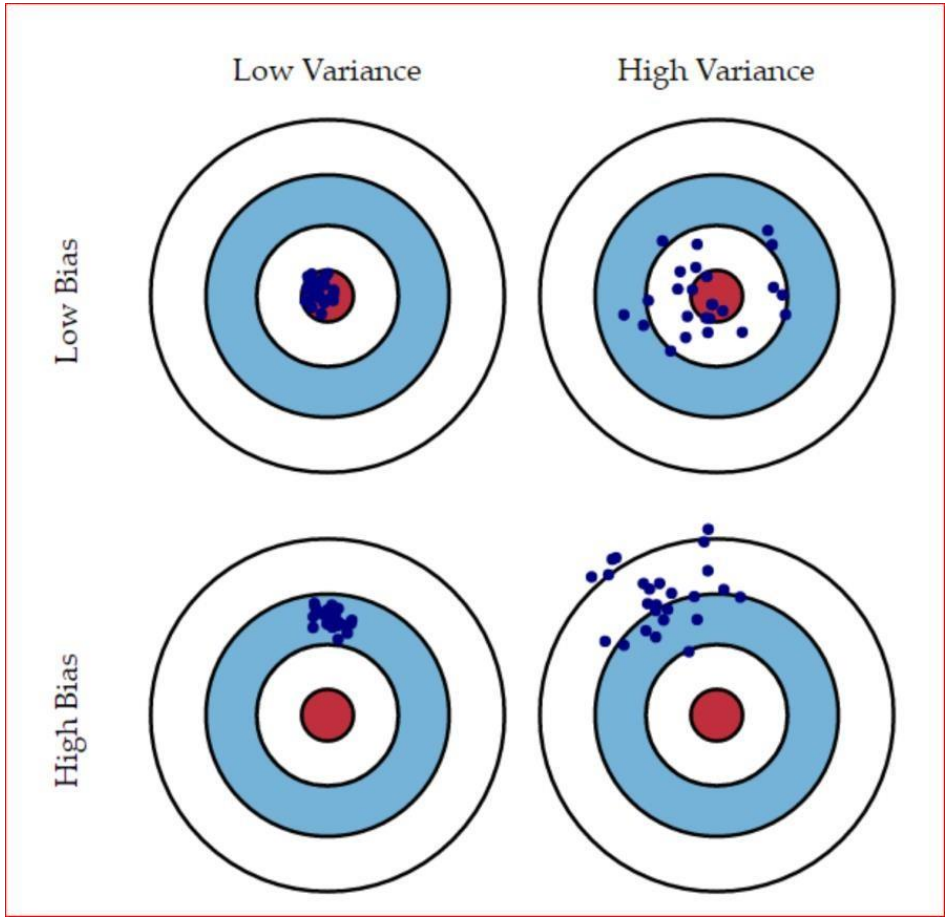
Bias and Variance

- **Variance** is the variability of model prediction for a given data point or a value which tells us spread of our data. Variance tells us how scattered are the predicted value from the actual value.
  - Model with high variance pays a lot of attention to training data and does not generalize on
  - the data which it hasn’t seen before.
  - As a result, such **models perform very well on training data but has high error rates on test data**.
  - High variance causes **overfitting** that implies that the algorithm models random noise
  - present in the training data.

Bias and Variance



Bias and Variance



Bias and Variance

- **Is there a way to find when we have a high bias or a high variance?**
  - High Bias can be identified when we have
    - High training error
    - Validation error or test error is close to training error
  - High Variance can be identified when
    - Low training error
    - High validation error or high test-error

Bias and Variance

- **How do we fix high bias or high variance in the data set?**
  - High bias is due to a simple model and we also see a high training error. To fix that we can do following things:
    - Add more input features
    - Add more complexity by introducing polynomial features
    - Decrease Regularization term
  - High variance is due to a model that tries to fit most of the training dataset points and hence gets more complex. To resolve high variance issue we need to work on
    - Getting more training data
    - Reduce input features
    - Increase Regularization term

Solutions

- Reduce the number of features
  - Manually select features
  - Model selection
- Regularization
  - Reduce magnitude/values of parameters θ_j.
  - Works well when we have a lot of features, each of which contributes a bit to the prediction.
- Bagging and Boosting

Sources

- Machine Learning, Andrew Ng, on Coursera by Stanford – a https://www.coursera.org/learn/machine-learning
- Deep Learning Specialization, Andrew Ng, on Coursera by deeplearning.ai
  - – https://www.coursera.org/specializations/deep-learning
- STATQUEST!!! An epic journey through statistics and machine learning, Josh Starmer, https://statquest.org/, https://www.youtube.com/channel/UCtYLUTtgS3k1Fg4y5tAhLbw
- https://towardsdatascience.com/cohens-kappa-9786ceceab58#:~:text=Cohen's%20kappa%20measures%20the%20agreement,raters%20may%20agree%20by%20chance.
- https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229
- https://repository.upenn.edu/cgi/viewcontent.cgi?article=1043&context=asc_papers#:~:text=Krippendorff's%20alpha%20(%CE%B1)%20is%20a,assign%20computable%2 0values%20to%20them.
- https://thenewstack.io/cohens-kappa-what-it-is-when-to-use-it-and-how-to-avoid-its-pitfalls/
- https://en.wikipedia.org/wiki/Cohen%27s_kappa