大数据分析技术

Chap. 10 Basic concept

王怡洋 副教授

大连海事大学 信息科学技术学院

内容提纲

注：该ppt请不要传播，☺。

10.1 What is data science?

- Collecting data from a wide variety of sources and putting them into a consistent format?
- Making observations about patterns in data?
- Visualizing trends in data?
- Making predictions about the tendency in the future?
- Identifying similarities between data points?
- Developing new machine learning and data mining algorithms?
- Accelerating algorithms?

10.2 Data analysis in “Practice”

Data of Vessel “*Pacific Vision*”

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	POSITION(LON)	POSITION(LAT)	COG	HDG	SOG	TRUE WIND(°)	TRUE WIND(m/s)	GUST(m/s)	WAVE(°)	WAVE(M)	SWELL(°)	SWELL(M)	SEAS(M)	CURRENT(°)	CURRENT(KN)	DISTANCE(NM)	TIME(HOURS)	SPEED(KN)	DRAUGHT(M)
2	029° 39.07W	08° 56.41S	128.6	126	10.8	291.9	6.04	6.65	111.0	0.68	94.5	1.68	1.8	180.0	0.23	1.91	0.17	11.22	23.1
3	029° 37.58W	08° 57.56S	128.3	127	10.8	291.9	6.04	6.65	111.0	0.68	94.5	1.68	1.8	180.0	0.23	1.88	0.17	11.04	23.1
4	029° 35.76W	08° 58.99S	128.0	127	10.8	291.9	6.04	6.4	114.0	0.6	96.0	1.72	1.8	180.0	0.23	2.29	0.21	10.92	23.1
5	029° 32.92W	09° 01.20S	129.2	127	10.7	291.9	6.04	6.4	114.0	0.6	96.0	1.72	1.8	198.4	0.25	3.57	0.33	10.83	23.1
6	029° 31.18W	09° 02.54S	127.2	127	10.8	291.9	6.04	6.4	114.0	0.6	96.0	1.72	1.8	198.4	0.25	2.19	0.2	10.93	23.1
7	029° 29.66W	09° 03.72S	127.1	127	10.6	290.9	6.31	6.7	114.0	0.6	96.0	1.72	1.8	198.4	0.25	1.91	0.17	11.23	23.1
8	029° 28.10W	09° 04.91S	127.7	127	10.7	290.9	6.31	6.7	114.0	0.6	96.0	1.72	1.8	213.7	0.28	1.95	0.18	10.83	23.1
9	029° 26.59W	09° 06.06S	127.8	127	10.7	290.9	6.31	6.7	114.0	0.6	96.0	1.72	1.8	213.7	0.28	1.88	0.17	11.08	23.1
10	029° 25.09W	09° 07.22S	128.0	127	10.7	290.9	6.31	6.7	114.0	0.6	96.0	1.72	1.8	213.7	0.28	1.88	0.17	11.07	23.1
11	029° 23.24W	09° 08.69S	128.9	127	10.6	288.7	6.23	6.5	108.0	0.48	97.5	1.72	1.8	198.4	0.25	2.35	0.22	10.67	23.1
12	029° 21.71W	09° 09.90S	128.9	127	10.5	288.6	6.44	6.75	109.5	0.48	99.0	1.72	1.8	198.4	0.25	1.94	0.18	10.78	23.1
13	029° 20.32W	09° 11.02S	128.9	127	10.5	288.6	6.44	6.75	109.5	0.48	99.0	1.72	1.8	198.4	0.25	1.77	0.16	11.06	23.1
14	029° 18.79W	09° 12.24S	129.2	127	10.5	288.6	6.44	6.75	109.5	0.48	99.0	1.72	1.8	198.4	0.25	1.94	0.18	10.78	23.1
15	029° 17.10W	09° 13.61S	129.5	126	10.6	288.6	6.44	6.75	109.5	0.48	99.0	1.72	1.8	198.4	0.25	2.16	0.2	10.81	23.1
16	029° 15.64W	09° 14.77S	129.4	127	10.7	288.6	6.44	6.75	109.5	0.48	99.0	1.72	1.8	198.4	0.25	1.85	0.17	10.91	23.1
17	029° 14.17W	09° 15.91S	127.0	125	10.6	288.6	6.44	6.75	109.5	0.48	99.0	1.72	1.8	198.4	0.25	1.84	0.17	10.84	23.1
18	029° 12.59W	09° 17.08S	126.6	125	10.7	287.9	6.67	7.0	103.5	0.4	94.5	1.7	1.74	270.0	0.31	1.95	0.18	10.85	23.1
19	029° 11.14W	09° 18.18S	126.8	125	10.6	287.9	6.67	7.0	103.5	0.4	94.5	1.7	1.74	270.0	0.31	1.81	0.16	11.32	23.1
20	029° 09.60W	09° 19.27S	125.1	123	10.6	287.9	6.67	7.0	103.5	0.4	94.5	1.7	1.74	270.0	0.31	1.87	0.17	11.02	23.1
21	029° 08.04W	09° 20.39S	125.9	123	10.7	287.9	6.67	7.0	103.5	0.4	94.5	1.7	1.74	219.8	0.3	1.9	0.17	11.2	23.1
22	029° 06.52W	09° 21.52S	127.3	125	10.7	288.7	6.39	7.0	109.5	0.5	94.5	1.7	1.74	219.8	0.3	1.88	0.17	11.08	23.1
23	029° 02.65W	09° 24.43S	127.4	124	10.7	288.0	6.47	7.15	112.5	0.36	94.5	1.7	1.74	219.8	0.3	4.8	0.44	10.91	23.1
24	029° 00.99W	09° 25.69S	127.1	125	10.7	288.0	6.47	7.15	112.5	0.36	94.5	1.7	1.74	219.8	0.3	2.07	0.19	10.91	23.1
25	028° 59.38W	09° 26.91S	128.3	126	10.7	288.0	6.47	7.15	112.5	0.36	94.5	1.7	1.74	219.8	0.3	2.0	0.18	11.12	23.1
26	028° 57.85W	09° 28.08S	127.6	126	10.7	288.0	6.47	7.15	112.5	0.36	94.5	1.7	1.74	219.8	0.3	1.91	0.17	11.23	23.1
27	028° 56.27W	09° 29.28S	127.1	126	10.6	288.6	6.91	7.35	112.5	0.36	94.5	1.7	1.74	219.8	0.3	1.97	0.18	10.92	23.1
28	028° 54.75W	09° 30.45S	128.5	127	10.6	288.6	6.91	7.35	112.5	0.36	94.5	1.7	1.74	219.8	0.3	1.91	0.17	11.24	23.1
29	028° 51.03W	09° 33.28S	127.2	127	10.4	289.2	6.67	7.35	118.5	0.36	96.0	1.7	1.74	219.8	0.3	4.64	0.43	10.79	23.1
30	028° 49.63W	09° 34.37S	127.8	128	10.5	289.2	6.67	7.35	118.5	0.36	96.0	1.7	1.74	219.8	0.3	1.76	0.16	10.99	23.1
31	028° 48.00W	09° 35.62S	127.9	128	10.4	289.7	6.69	7.15	124.5	0.7	82.5	1.58	1.7	213.7	0.28	2.04	0.19	10.72	23.1
32	028° 46.41W	09° 36.82S	127.5	128	10.4	289.7	6.69	7.15	124.5	0.7	82.5	1.58	1.7	213.7	0.28	1.98	0.18	10.99	23.1
33	028° 44.92W	09° 37.95S	127.7	128	10.5	289.7	6.69	6.95	126.0	0.6	84.0	1.58	1.7	213.7	0.28	1.86	0.17	10.93	23.1
34	028° 43.46W	09° 39.06S	127.3	128	10.4	289.7	6.69	6.95	126.0	0.6	84.0	1.58	1.7	213.7	0.28	1.82	0.17	10.7	23.1
35	028° 42.01W	09° 40.18S	127.9	128	10.5	289.7	6.69	6.95	126.0	0.6	84.0	1.58	1.7	213.7	0.28	1.82	0.17	10.69	23.1
36	028° 40.57W	09° 41.28S	127.1	128	10.6	289.7	6.69	6.95	126.0	0.6	84.0	1.58	1.7	213.7	0.28	1.8	0.16	11.23	23.1
37	028° 39.06W	09° 42.41S	127.2	128	10.6	289.8	6.64	7.15	126.0	0.6	84.0	1.58	1.7	213.7	0.28	1.87	0.17	11.0	23.1
38	028° 37.57W	09° 43.53S	127.3	128	10.6	289.8	6.64	7.15	126.0	0.6	84.0	1.58	1.7	213.7	0.28	1.85	0.17	10.85	23.1
39	028° 35.95W	09° 44.75S	127.3	128	10.6	289.8	6.64	7.15	127.5	0.6	87.0	1.58	1.7	213.7	0.28	2.01	0.18	11.19	23.1
40	028° 34.45W	09° 45.90S	128.2	129	10.6	289.8	6.64	7.15	127.5	0.6	87.0	1.58	1.7	213.7	0.28	1.87	0.17	11.01	23.1
41	028° 32.99W	09° 47.04S	128.1	128	10.6	289.8	6.64	7.15	127.5	0.6	87.0	1.58	1.7	213.7	0.28	1.83	0.17	10.78	23.1
42	028° 31.09W	09° 48.52S	128.5	129	10.7	287.8	7.19	7.5	127.5	0.6	87.0	1.58	1.7	213.7	0.28	2.39	0.22	10.86	23.1
43	028° 29.51W	09° 49.76S	128.6	129	10.7	287.8	7.19	7.5	127.5	0.6	87.0	1.58	1.7	213.7	0.28	2.0	0.18	11.09	23.1
44	028° 28.00W	09° 50.94S	128.5	129	10.8	287.8	7.19	7.5	127.5	0.6	87.0	1.58	1.7	213.7	0.28	1.9	0.17	11.17	23.1
45	028° 26.37W	09° 52.19S	128.0	128	10.8	287.8	7.19	7.5	127.5	0.6	87.0	1.58	1.7	213.7	0.28	2.04	0.18	11.34	23.1
46	028° 24.64W	09° 53.49S	127.6	128	10.8	288.6	6.91	7.35	127.5	0.5	87.0	1.58	1.64	213.7	0.28	2.14	0.19	11.28	23.1
47	028° 22.29W	09° 55.29S	128.0	128	10.7	290.6	7.1	7.4	120.0	0.76	85.5	1.54	1.68	239.0	0.23	2.94	0.27	10.89	23.1
48	028° 20.63W	09° 56.56S	128.1	128	10.7	290.6	7.1	7.4	120.0	0.76	85.5	1.54	1.68	239.0	0.23	2.06	0.19	10.86	23.1

... 2019-05 | 2019-06 | 2019-07 | 2019-08 | 2019-09 | 2019-10 | 2019-11 | 2019-12 | 2020-01 | 2020-02 | 2020-03 | 2020-04 | 2020-05 | 2020-06 | 2020-07 | 2020-08

- What might we want to learn about them?

10.2.1 *Descriptive Statistics*

- What is the mean SOG of these data of *the whole/each voyage/each day*?
- How far are the data (whole, voyage, day) spread out from their average mean?
- What is the distribution of each metocean factor (whole, voyage, day)?
- Does there exist correlation among these variables?
- …

10.2.2 *Reasoning about data*

- How does the mean SOG (whole, voyage, day) compare to the service speed? Is it higher or lower?

- Does there exists relation with the metocean conditions?

- Where does the variance of data come from?

- Anything to say about the metocean distribution?

- If there exists a correlation between variables, does it make sense?

- …

10.2.3 *Making predictions*

- Can we make a good prediction of SOG?
- How do we choose which predictor (method) to use?
- How do we measure the performance of the method? (statistic measurements, …)
- If there contains a train phase, how do we train? (data, parameter tuning, …)
- Do we need to put all the metocean factors as the input?
  Difference between using more or less input features?
  Can we tell the best combination of the inputs?
- …

10.2.4 *Anomaly detection*

- Does there exists any anomaly in the data (whole, voyage, day)?
- What caused these anomalies?
- How can we recognize them automatically?
- Do they affect the results of predictors?
  Do we need to remove them from the training phase?
- How can we develop a more reliable model?
- …

Data: $\{x_1, x_2, \cdots, x_n\}; \{y_1, y_2, \cdots, y_n\}$

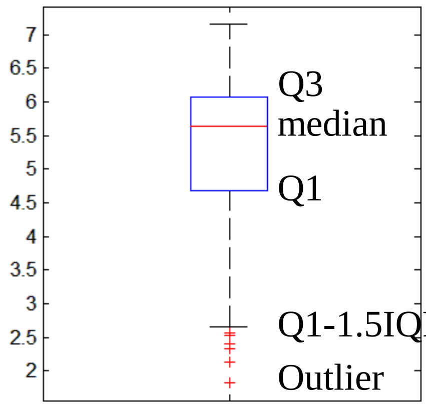
- Average mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Variance (standard deviation)

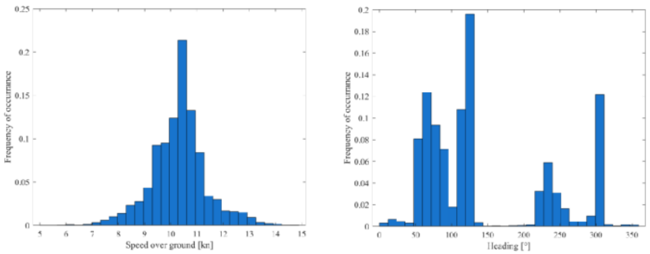
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Box plot



IQR (interquartile range)
= Q3-Q1

- Distribution

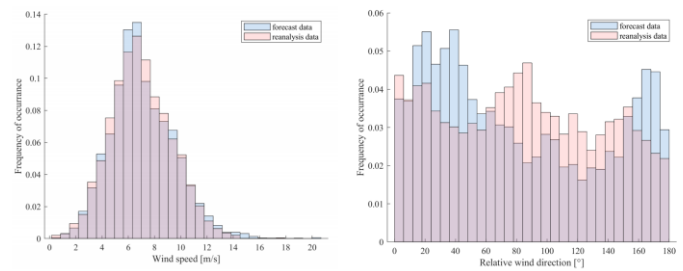


- (Pearson) Correlation coefficient

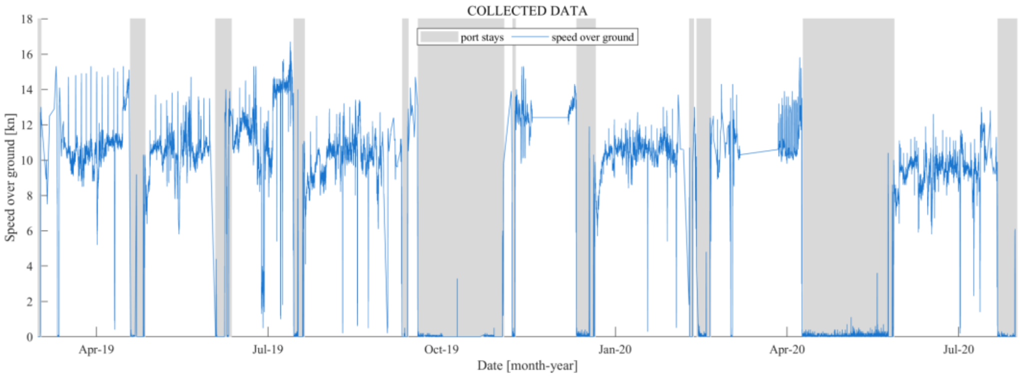
$$\rho_{xy} = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

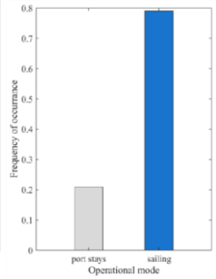
Notice:

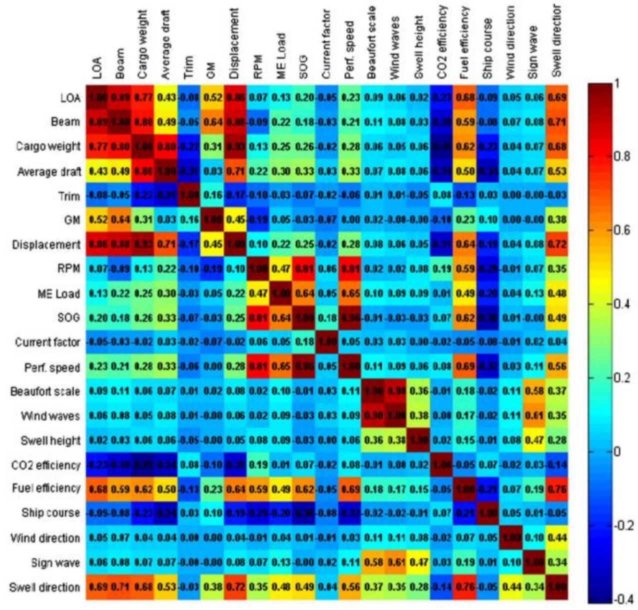
- linear correlation: it cannot capture nonlinear relationships between two variables
- $[-1, 1]$

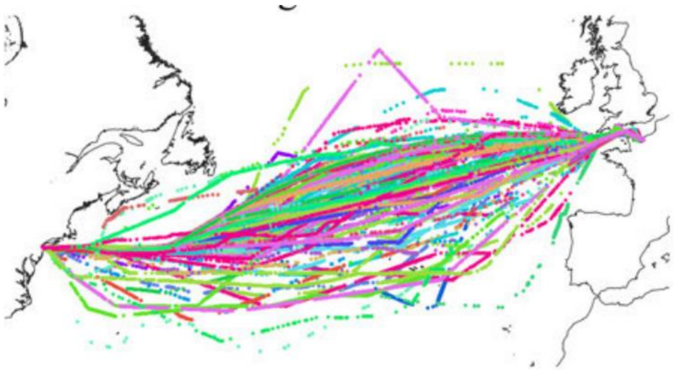


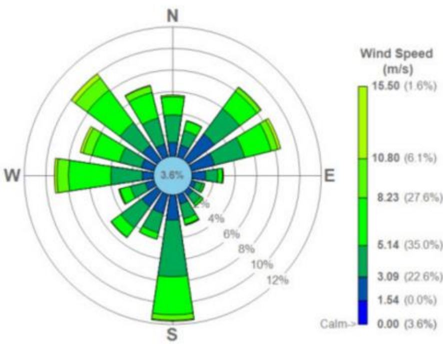
10.3 Descriptive statistics











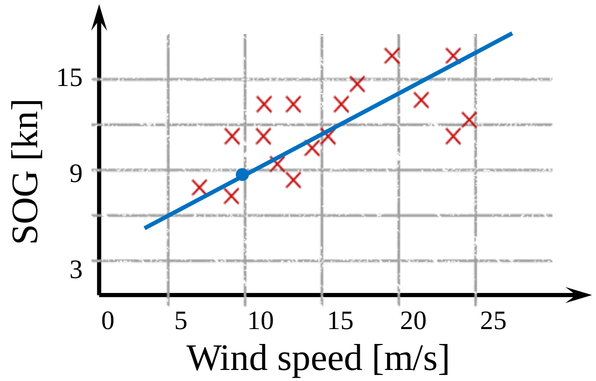
On data **On prediction** **On error**

Error = Prediction – ground truth (data)

10.4 First look of linear regression

NAME	NOTES
SOG (kn)	Speed over ground，对地速度
DRAUGHT (m)	吃水
COG (°)	Course over ground，对地航向
HDG (°)	Heading，船艏向
CURRENT (°)	流向
TRUE WIND (°)	风向
WAVE (°)	风浪方向
SWELL (°)	涌浪方向
CURRENT (kn)	流速
TRUE WIND (m/s)	风速
WAVE (m)	浪高
SWELL (m)	涌浪高度
GUST (m/s)	阵风风速
SEAS (m)	耦合浪高

- Wind speed → SOG



- *Supervised learning*: Give the “right answer” for each example in the data;

- *Regression* problem: Predict real-valued output
  (Vs. *Classification* problem)

10.4.1 Dataset notations

- Wind speed → SOG

TRUE WIND(m/s)	SOG
6.04	10.8
6.04	10.8
6.04	10.8
6.04	10.7
6.04	10.8
6.31	10.6
6.31	10.7
6.31	10.7
6.31	10.7
6.23	10.6
6.44	10.5
6.44	10.5
6.44	10.5
6.44	10.6
6.44	10.7
6.44	10.6
6.67	10.7
6.67	10.6
6.67	10.6
6.67	10.7
6.39	10.7
6.47	10.7
6.47	10.7
6.47	10.7
6.47	10.7
6.91	10.6
6.91	10.6
6.67	10.4
6.67	10.5
6.69	10.4
6.69	10.4
6.69	10.5
6.69	10.4
6.69	10.5
6.69	10.6
6.64	10.6
6.64	10.6
6.64	10.6

- Dataset (Training set)

- Notations:

m: the number of training examples

$\mathbf{x}$: "input" variable/feature (wind speed) $\mathbf{x} = (x_1, x_2, \cdots, x_m)^\mathsf{T}$

$\mathbf{y}$: ground truth/label (SOG) $\mathbf{y} = (y_1, y_2, \cdots, y_m)^\mathsf{T}$

$\mathbf{z}$: "output" variable/target (prediction) $\mathbf{z} = (z_1, z_2, \cdots, z_m)^\mathsf{T}$

$\{(x_i, y_i)\}_{i=1,\cdots,m}$: the i^{th} training example

10.4.2 Our goal

- Notations:

m: the number of training examples

$\mathbf{x}$: "input" variable/feature (wind speed)

$\mathbf{y}$: ground truth/label (SOG)

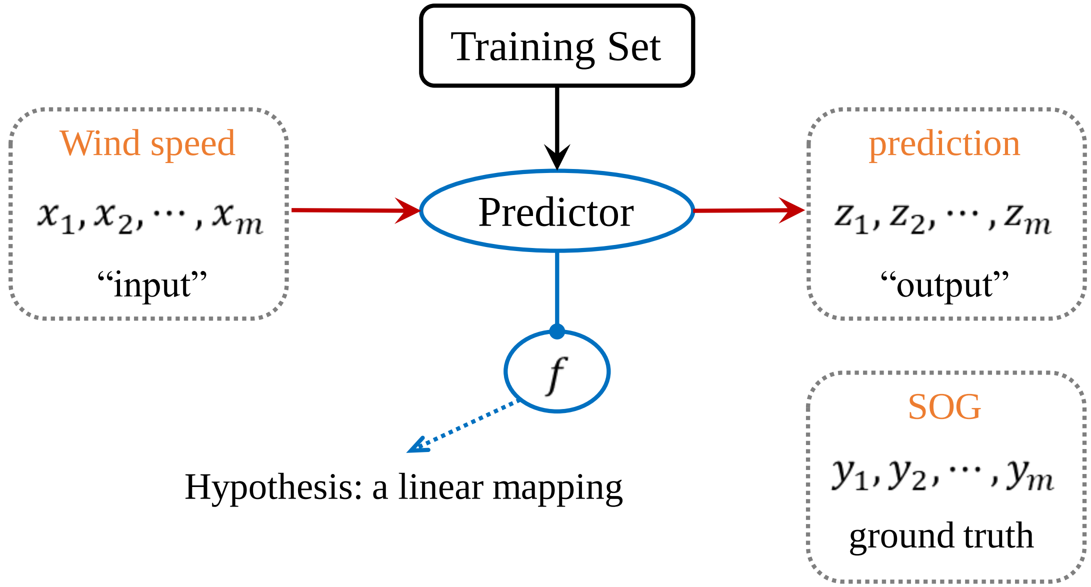
$\mathbf{z}$: "output" variable/target (prediction)

$$\mathbf{x} = (x_1, x_2, \cdots, x_m)^{\mathsf{T}}$$

$$\mathbf{y} = (y_1, y_2, \cdots, y_m)^{\mathsf{T}}$$

$$\mathbf{z} = (z_1, z_2, \cdots, z_m)^{\mathsf{T}}$$

$\{(x_i, y_i)\}_{i=1,\cdots,m}$: the i$^{\text{th}}$ training example



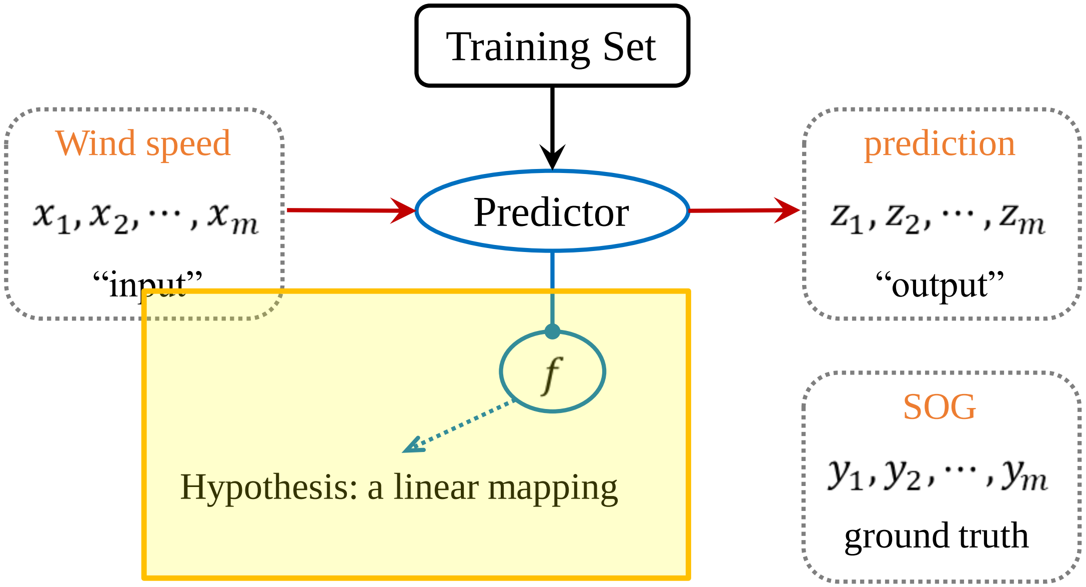
Goal: predict SOG by wind speed

$$f(x_i) = z_i \approx\rightarrow y_i$$

$$\varepsilon_i = z_i - y_i$$

A good predictor f means: ε_i as small as possible

10.4.3 How do we present a linear predictor



Goal: predict SOG by wind speed

$$f(x_i) = z_i \approx\rightarrow y_i$$

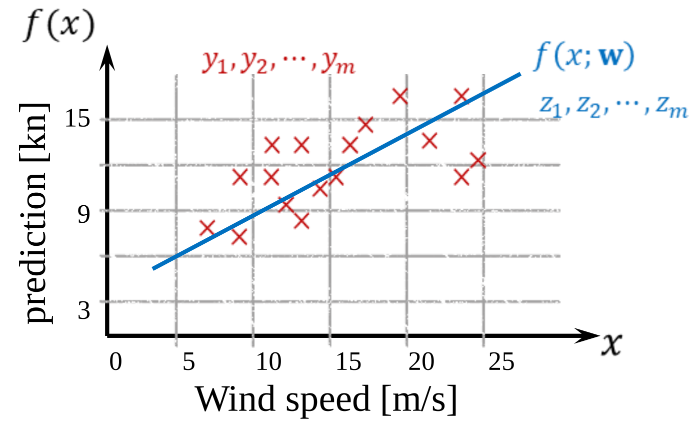
$$\varepsilon_i = z_i - y_i$$

A good predictor f means: ε_i as small as possible

How do we present f?

$$f(x; \mathbf{w}) = w_o + w_1 x$$

$$\mathbf{w} = (w_0, w_1)^\intercal$$

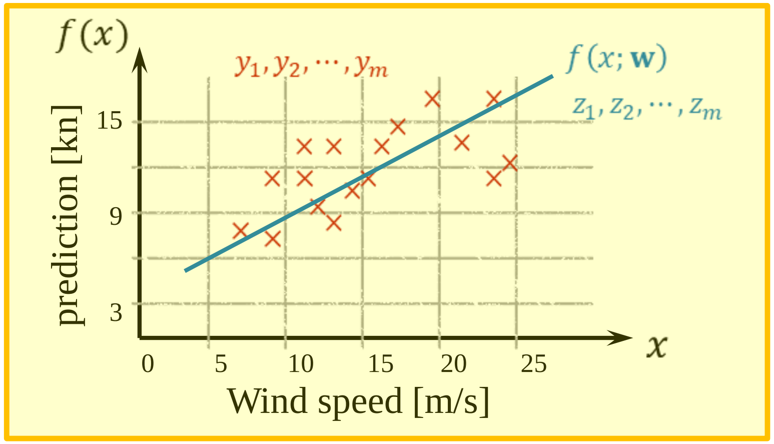


- Linear regression with one variable
- *Univariate linear regression*

10.4.4 How to choose the parameter

- *Univariate linear regression*

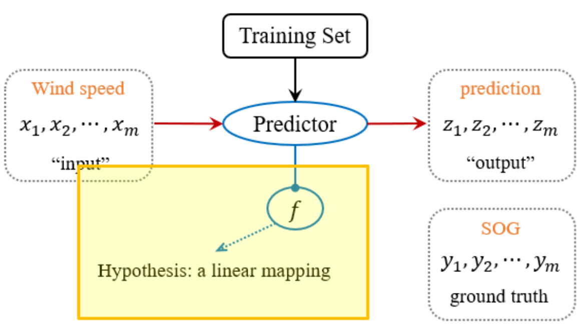
$$f(x;\mathbf{w}) = w_o + w_1 x \qquad \mathbf{w} = (w_0, w_1)^{\intercal}$$



- Goal: predict SOG by wind speed

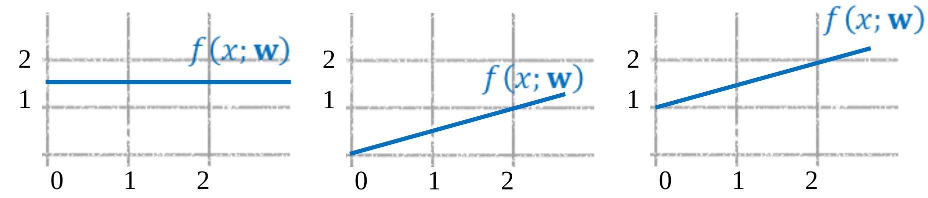
$$f(x_i;\mathbf{w}) = z_i \approx\rightarrow y_i \qquad \varepsilon_i = z_i - y_i$$

- A good predictor f means: ε_i as small as possible



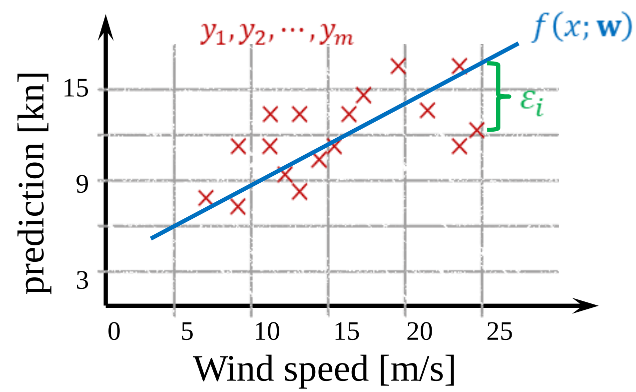
- How to choose the parameter $\mathbf{w}$?

$$f(x;\mathbf{w}) = w_o + w_1 x$$



$w_o = 1.5$	$w_o = 0$	$w_o = 1$
$w_1 = 0$	$w_1 = 0.5$	$w_1 = 0.5$

- Choose w_o and w_1 so that $f(x;\mathbf{w})$ **is close to** the ground truth for our training examples

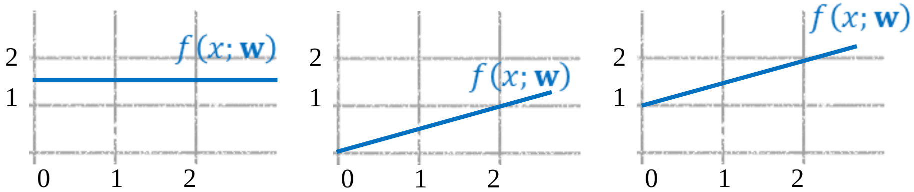


- ε_i as small as possible

10.4.5 How do we present a linear predictor

- How to choose the parameter $\mathbf{w}$?

$$f(x;\mathbf{w}) = w_o + w_1 x$$

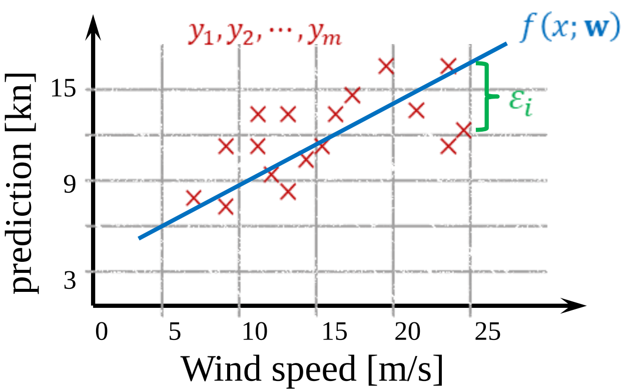


$w_o = 1.5$, $w_1 = 0$

$w_o = 0$, $w_1 = 0.5$

$w_o = 1$, $w_1 = 0.5$

- Choose w_o and w_1 so that $f(x;\mathbf{w})$ **is close to** the ground truth for our training examples



- ε_i as small as possible

$$\min_{w_o,w_1} (f(x_i;\mathbf{w}) - y_i)^2$$

$$\min_{\mathbf{w}} \sum_{i=1}^{m} (f(x_i;\mathbf{w}) - y_i)^2$$

$$\min_{w_o,w_1} \frac{1}{2m} \sum_{i=1}^{m} (w_o + w_1 x_i - y_i)^2$$

Let's define: $\mathcal{L}(w_o, w_1) = \frac{1}{2m}\sum_{i=1}^{m}(w_o + w_1 x_i - y_i)^2$

$$\min_{w_o,w_1} \mathcal{L}(w_o, w_1)$$

loss/cost function (squared error function)

- Different hypothesis on f, different $\mathcal{L}$

- Different w_o and w_1, different value of $\mathcal{L}$.
- The best w_o and w_1 are corresponding to the lowest value of $\mathcal{L}$, which are defined as $\mathbf{w}^*$, i.e.

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w})$$

- $\mathcal{L}(w_0^*, w_1^*)$ is the best predictor under the linear hypothesis

10.4.6 How to get the best parameter

$$\min_{w_o,w_1} (f(x_i;\mathbf{w}) - y_i)^2$$

$$\min_{\mathbf{w}} \sum_{i=1}^{m} (f(x_i;\mathbf{w}) - y_i)^2$$

||

$$\min_{w_o,w_1} \frac{1}{2m} \sum_{i=1}^{m} (w_o + w_1 x_i - y_i)^2$$

Let's define: $\mathcal{L}(w_o, w_1) = \frac{1}{2m}\sum_{i=1}^{m}(w_o + w_1 x_i - y_i)^2$

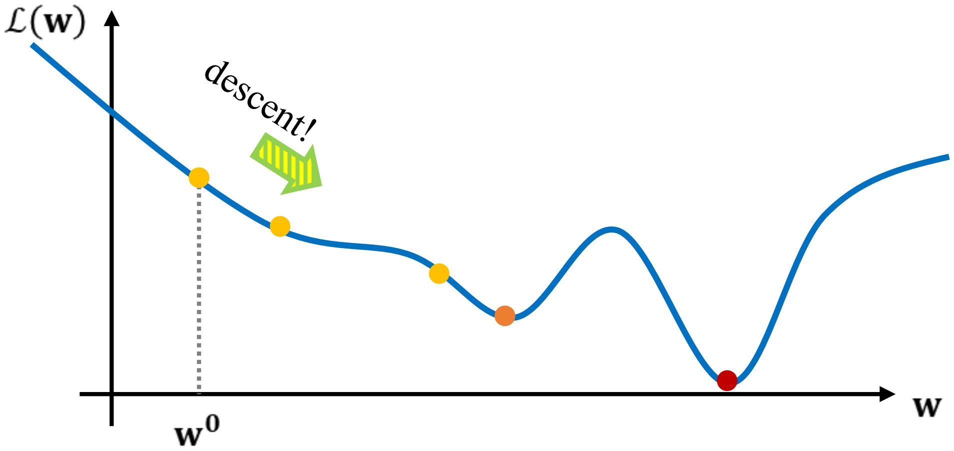
$$\min_{w_o,w_1} \mathcal{L}(w_o, w_1)$$

loss/cost function (squared error function)

- Different hypothesis on f, different $\mathcal{L}$
- Different w_o and w_1, different value of $\mathcal{L}$.
- The best w_o and w_1 are corresponding to the lowest value of $\mathcal{L}$, which are defined as $\mathbf{w}^*$, i.e.

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w})$$

- $\mathcal{L}(w_0^*, w_1^*)$ is the best predictor under the linear hypothesis



- Goal: $\mathcal{L}(\mathbf{w}^1) \ll \mathcal{L}(\mathbf{w}^0)$

From Taylor expansion:

$$\mathcal{L}(\mathbf{w}^1)\text{-}\mathcal{L}(\mathbf{w}^0) = (\mathbf{w}^1 - \mathbf{w}^0)\nabla\mathcal{L}(\mathbf{w}^0)$$

- $\mathcal{L}(\mathbf{w}^1)\text{-}\mathcal{L}(\mathbf{w}^0) < 0$
- For faster descent: $\mathcal{L}(\mathbf{w}^1) - \mathcal{L}(\mathbf{w}^0)$ as small as possible

$$(\mathbf{w}^1 - \mathbf{w}^0) = -\nabla\mathcal{L}(\mathbf{w}^0)$$

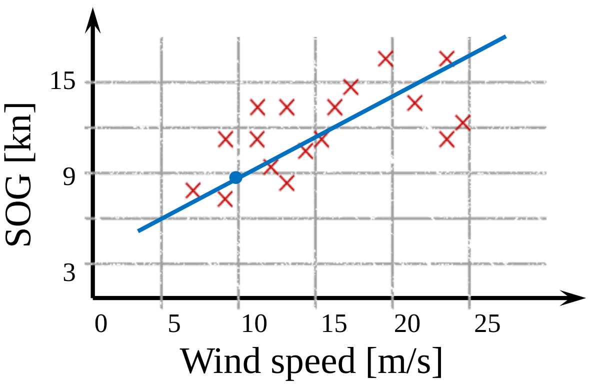
$$\mathbf{w}^1 = \mathbf{w}^0 - \nabla\mathcal{L}(\mathbf{w}^0)$$

$\boxed{\text{Gradient descent Alg.:} \quad \mathbf{w}^{t+1} = \mathbf{w}^t - \alpha\nabla\mathcal{L}(\mathbf{w}^t)}$

Learning rate

10.4.7 A review of linear regression

- Wind speed → SOG



- A good predictor f means: ε_i as small as possible

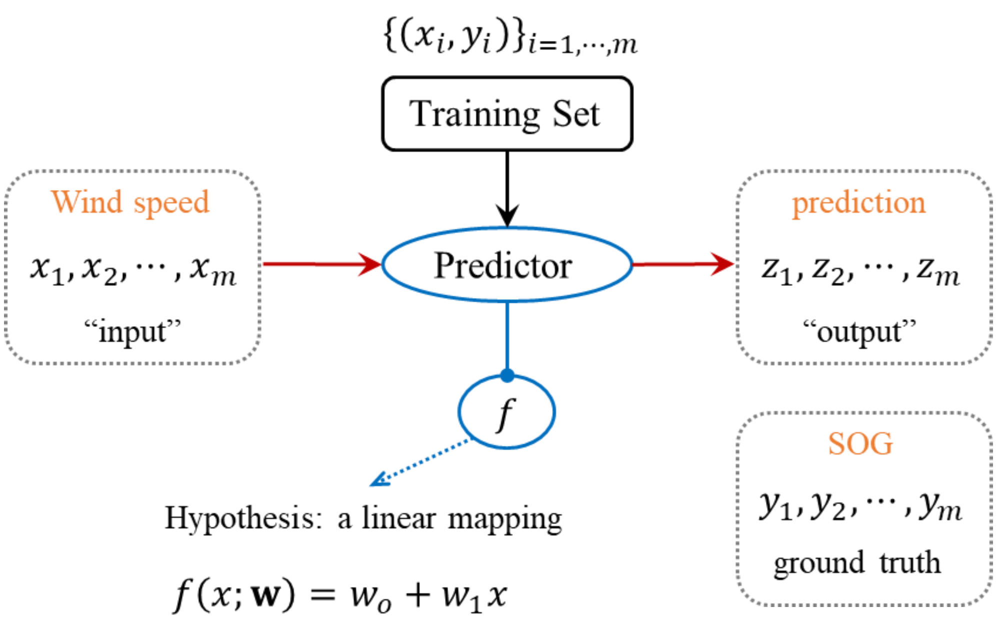
$$\min_{\mathbf{w}} \sum_{i=1}^{m} \left(f(x_i;\mathbf{w}) - y_i\right)^2$$

Let's define: $\mathcal{L}(w_o, w_1) = \frac{1}{2m}\sum_{i=1}^{m}(w_o + w_1 x_i - y_i)^2$

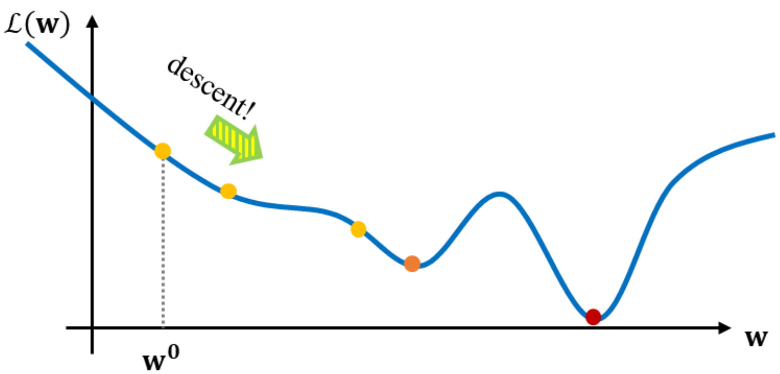
$$\min_{w_o, w_1} \mathcal{L}(w_o, w_1)$$

- The best w_o and w_1 are corresponding to the lowest value of $\mathcal{L}$, which are defined as $\mathbf{w}^*$, i.e.

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w})$$

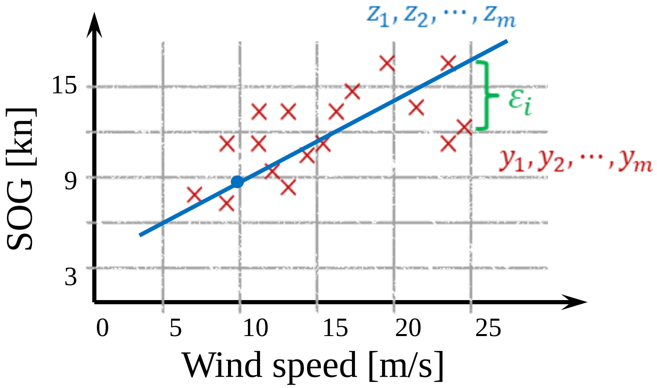


$$f(x;\mathbf{w}) = w_o + w_1 x$$



Gradient descent Alg.: $\mathbf{w}^{t+1} = \mathbf{w}^t - \alpha \nabla \mathcal{L}(\mathbf{w}^t)$

10.5 How's the results?



$$z_i = f(x_i;\ \mathbf{w}^*)$$

$$\varepsilon_i = f(x_i;\ \mathbf{w}^*) - y_i$$

- Average error on training data:

$$\sum_{i=1}^{m} \varepsilon_i = \sum_{i=1}^{m} (f(x_i;\ \mathbf{w}^*) - y_i)$$

- However, we care less on the error on training data ☹
- What we really care about is the error on new data (testing data)!

	dataset	function	prediction	error
training stage	$\{x_i, y_i\}_{i=1,\cdots,m}$	$f(x;\mathbf{w}^*)$	$\{z_i\}_{i=1,\cdots,m}$	$\{\varepsilon_i\}_{i=1,\cdots,m}$
testing stage	$\{\tilde{x}_i, \tilde{y}_i\}_{i=1,\cdots,n}$	$f(x;\mathbf{w}^*)$	$\{\tilde{z}_i\}_{i=1,\cdots,m}$	$\{\tilde{\varepsilon}_i\}_{i=1,\cdots,m}$

- Use another hypothesis to create a new function: $f(x;\mathbf{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5$
- Then we have different error (average error) for each function

10.5 How's the results?

- Use another hypothesis to create a new function: $f(x;\mathbf{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5$
- Then we have different error (average error) for each function

---

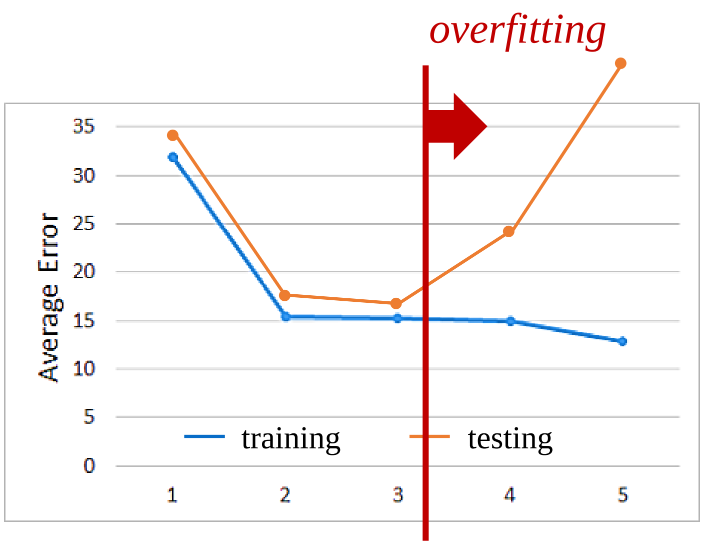
- How's the error on training data?
- Why?

- A more complex model yields lower error on training data ☺
  (If we can truly find the best function)

- But how's the results on testing data?
- A more complex model does not always lead to better performance on testing data ☹

- That is called *overfitting* !
- So, choose an appropriate model ☺

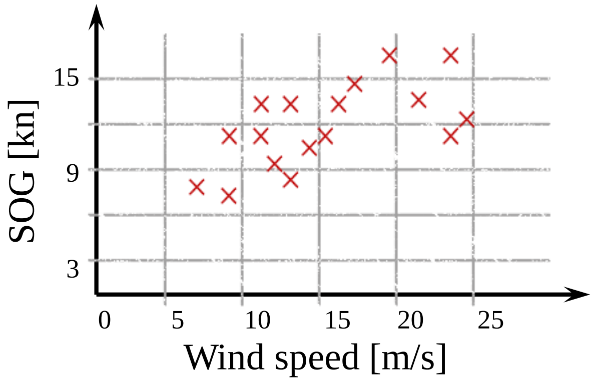
- (Taking a driving license exam…)

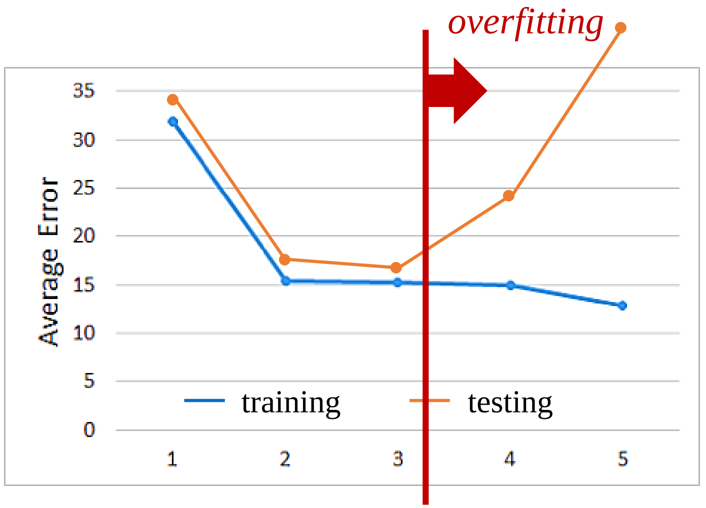


FOR THE FINAL TIME - YES I DID NOTICE HOW WELL YOU OPENED THE DOOR WHEN YOU GOT IN

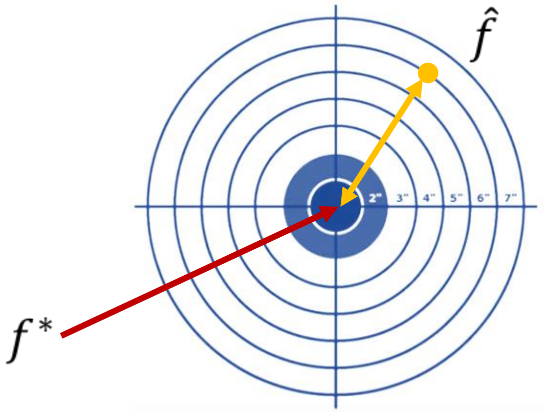
10.6 Where does the error come from?

- Wind speed → SOG





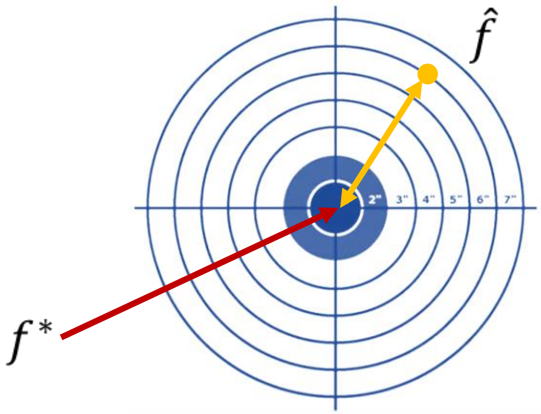
- In theory, there exists an optimal function f^* such that $f^*(x_i) = y_i$, but we don't know its exact form
- From training data, we try to find an $\hat{f}$ to approach f^*, i.e., $\hat{f}$ is an estimator of f^*



- For a given training dataset, suppose we can get the optimal $\hat{f}$

- Bias + variance
- What's the meaning of bias and variance of an estimator?

10.6.1 Bias and variance of an estimator



- For a given training dataset, suppose we can get the optimal $\hat{f}$
- Bias + variance
- What's the meaning of bias and variance of an estimator?

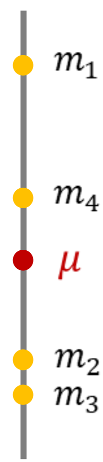
- Estimate the mean of a variable x
  - assume the mean of x is μ
  - assume the variance of x is σ^2

- Estimate the mean μ
  - sample N points: $x_1, x_2, \cdots, x_N$
  - The mean of N samples is:

$$m = \frac{1}{N}\sum_{i=1}^{N} x_i \neq \mu$$

$$E[m] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} E[x_i] = \mu$$

  - So, the estimator m is unbiased ☺

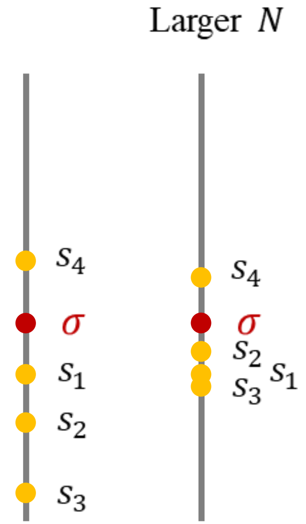


- Estimate the variance σ^2
  - The variance of N samples is:

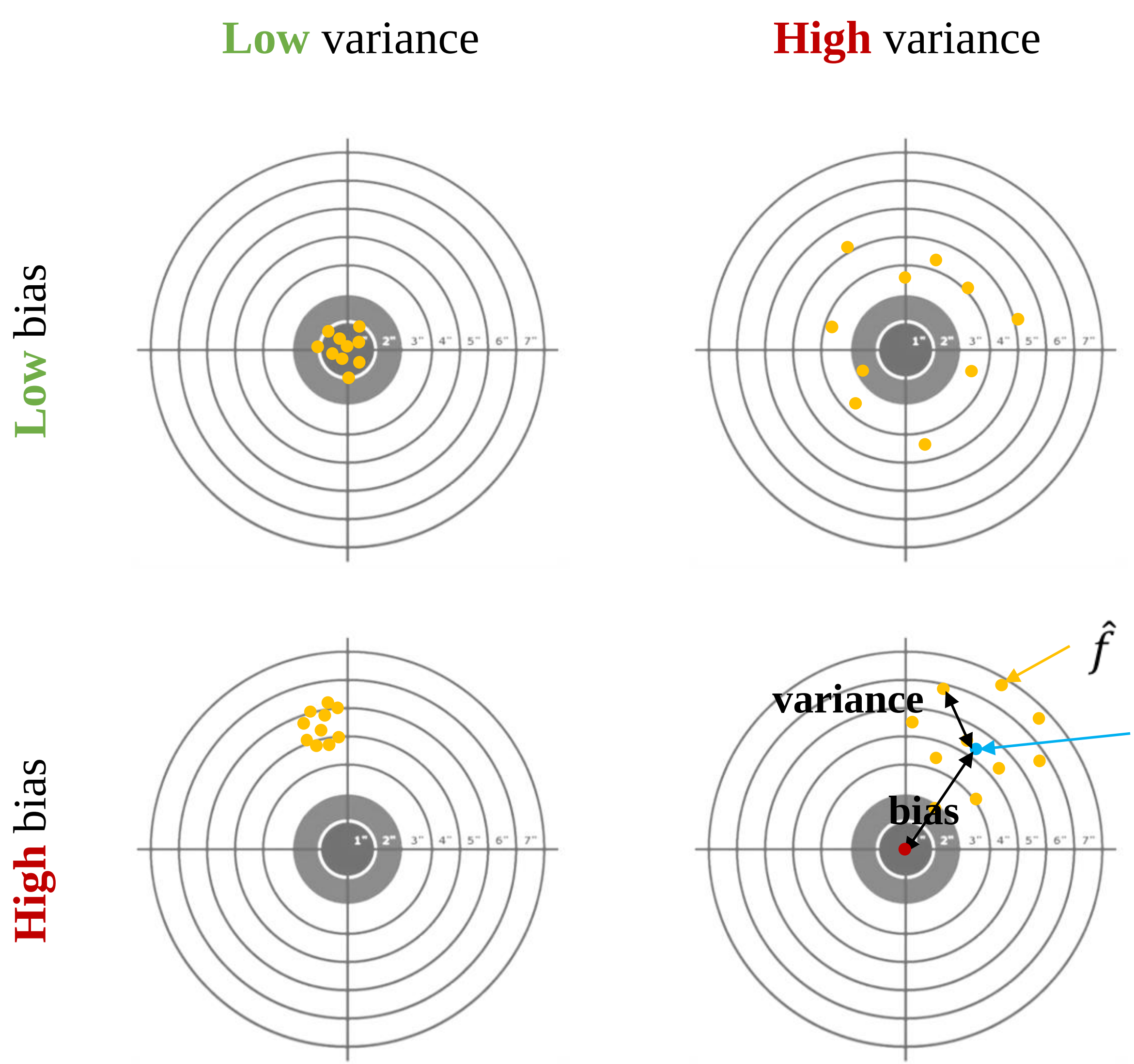
$$s^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - m)^2$$

$$E[s^2] = \frac{N-1}{N}\sigma^2 \neq \sigma^2$$

  - the estimator s^2 is biased ☹

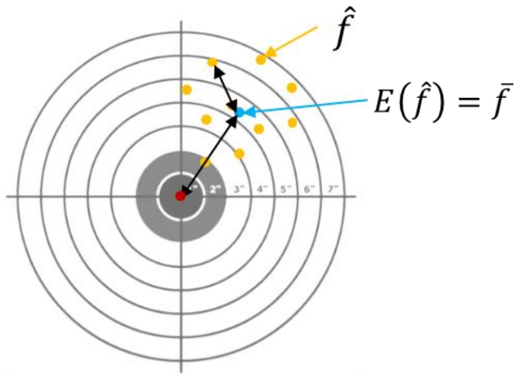


10.6.1 Bias and variance of an estimator



- For a given training set, we get one $\hat{f}$
- So, the error comes from ***both the bias and variance***.

10.6.1 Bias and variance of an estimator



- For a given dataset $\mathcal{D}$, let $f(x;\ \mathcal{D})$ be the well-trained model on $\mathcal{D}$.
- So x is the input feature, $y_{\mathcal{D}}$ is the label of $\mathcal{D}$, and y is the ground true data of x.
- Then for different dataset, we have an expected prediction of f, i.e.,

$$\bar{f}(x) = E_{\mathcal{D}}[f(x;\ \mathcal{D})]$$

- And its variance is:

$$var(x) = E_{\mathcal{D}}\left[\left(f(x;\ \mathcal{D}) - \bar{f}(x)\right)^2\right]$$

- The noise is:

$$\varepsilon^2 = E_{\mathcal{D}}[(y_{\mathcal{D}} - y)^2]$$

and we suppose:

$$E_{\mathcal{D}}[y_{\mathcal{D}} - y] = 0$$

- The bias from prediction and the ground truth is:

$$bias^2 = \left(\bar{f}(x) - y\right)^2$$

- So let's see the error of predictor $f(x;\mathcal{D})$ on $\mathcal{D}$:

$$E_{\mathcal{D}}[(f(x;\ \mathcal{D}) - y_{\mathcal{D}})^2]$$

$$= E_{\mathcal{D}}\left[\left(f(x;\ \mathcal{D}) - \bar{f}(x) + \bar{f}(x) - y_{\mathcal{D}}\right)^2\right]$$

$$= E_{\mathcal{D}}\left[\left(f(x;\ \mathcal{D}) - \bar{f}(x)\right)^2\right] + E_{\mathcal{D}}\left[\left(\bar{f}(x) - y_{\mathcal{D}}\right)^2\right]$$

$$= E_{\mathcal{D}}\left[\left(f(x;\ \mathcal{D}) - \bar{f}(x)\right)^2\right] + E_{\mathcal{D}}\left[\left(\bar{f}(x) - y + y - y_{\mathcal{D}}\right)^2\right]$$

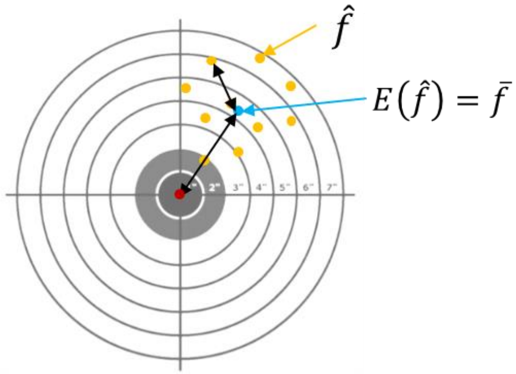
$$= E_{\mathcal{D}}\left[\left(f(x;\ \mathcal{D}) - \bar{f}(x)\right)^2\right] + E_{\mathcal{D}}\left[\left(\bar{f}(x) - y\right)^2\right] + E_{\mathcal{D}}[(y - y_{\mathcal{D}})^2]$$

$$+2E_{\mathcal{D}}[(\bar{f}(x) - y)(y - y_{\mathcal{D}})]$$

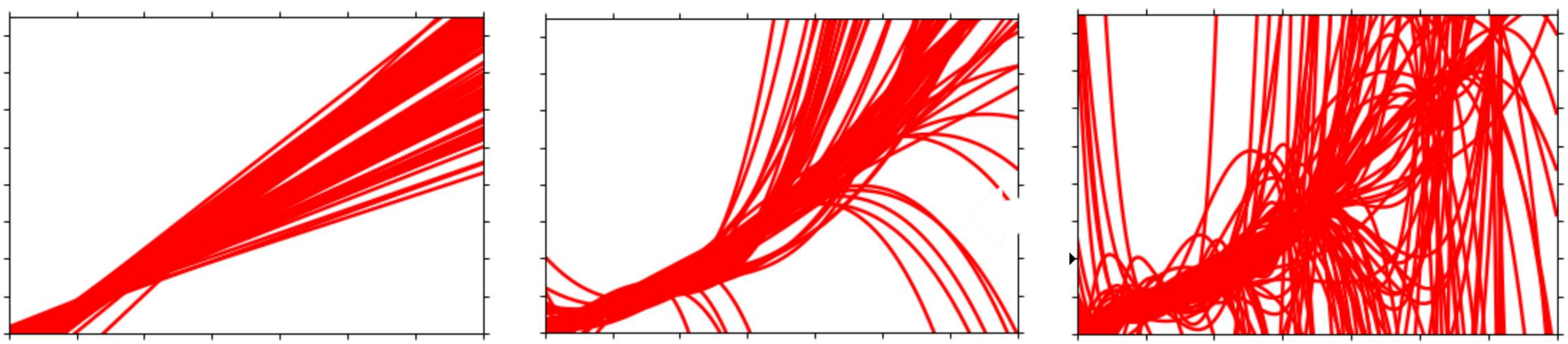
$$= E_{\mathcal{D}}\left[\left(f(x;\ \mathcal{D}) - \bar{f}(x)\right)^2\right] + \left(\bar{f}(x) - y\right)^2 + E_{\mathcal{D}}[(y - y_{\mathcal{D}})^2]$$

$$= bias^2 + var(x) + \varepsilon^2$$

10.6.2 Parallel universes



- In each universe, “I” kept recording the data in a regular cycle, i.e, different universes will have different training data
- In different universes, we use the same model, but obtain different $\hat{f}$
- $\hat{f}$ in 100 “universes”?



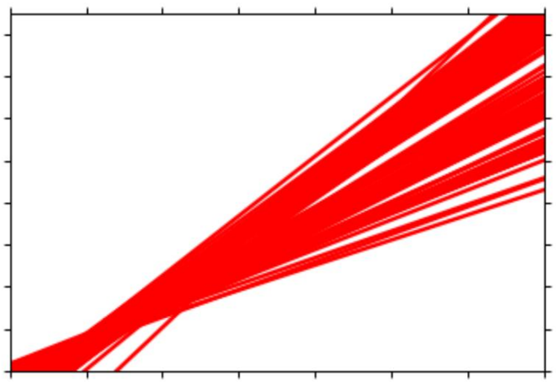
$f(x; \mathbf{w}) = w_0 + w_1 x$

$f(x; \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$

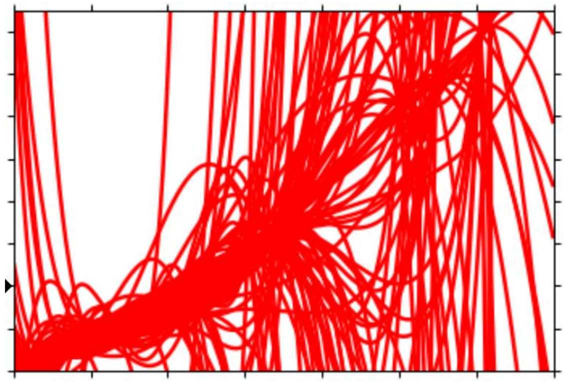
$f(x; \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5$

Figures from Hongyi Li

10.6.3 Variance



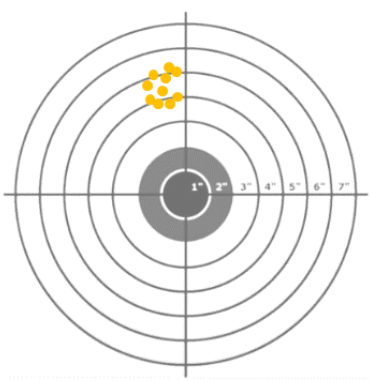
$f(x; \mathbf{w}) = w_0 + w_1 x$

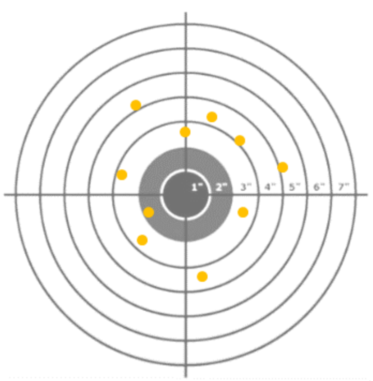


$$f(x; \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5$$

small variance

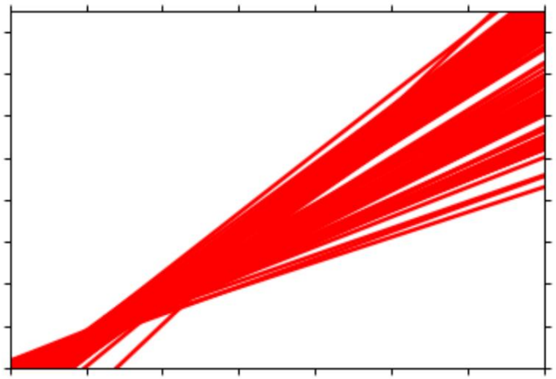
large variance



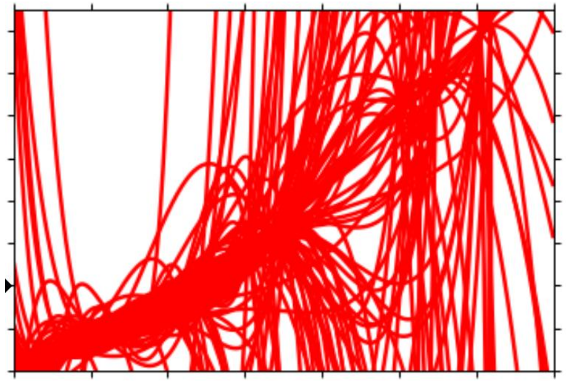


- ***Simpler model is less influenced by the sampled data***
- Consider the extreme case: $f(x) = c$

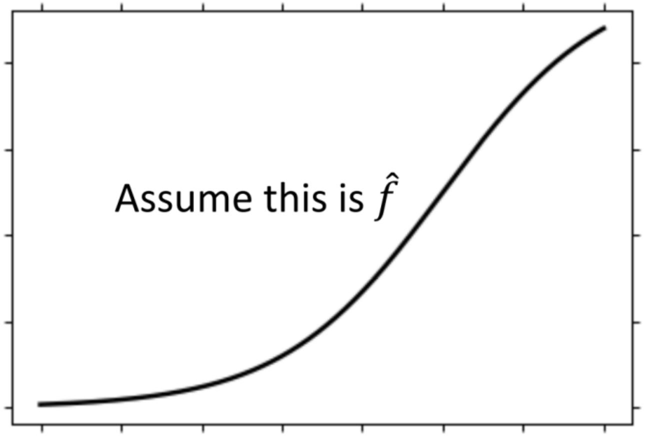
10.6.4 Bias



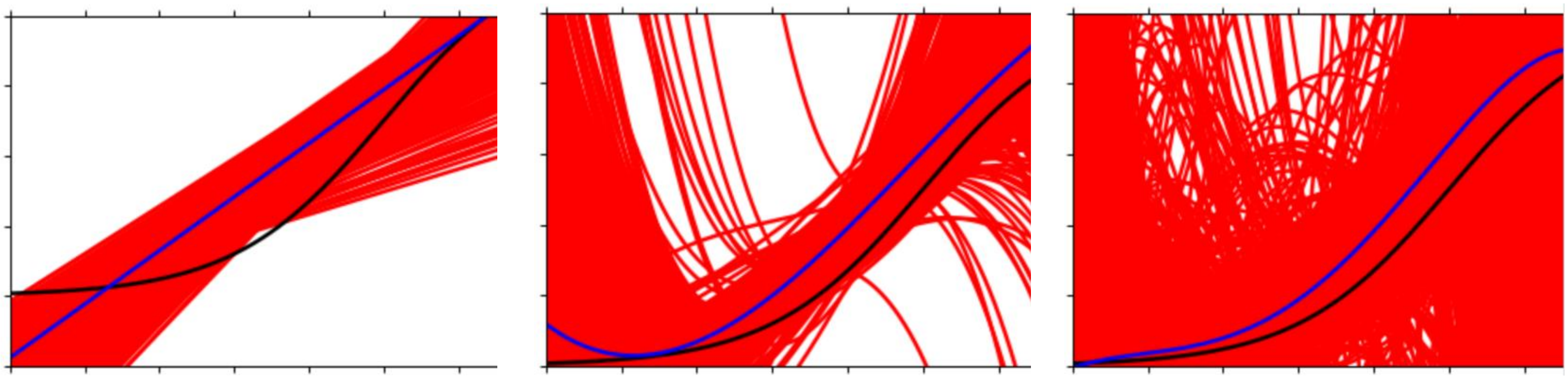
$f(x;\mathbf{w}) = w_0 + w_1 x$



$f(x;\mathbf{w}) = w_0 + w_1 x + w_2 x^2$
$+ w_3 x^3 + w_4 x^4 + w_5 x^5$



10.6.4 Bias



$f(x;\mathbf{w}) = w_0 + w_1 x$

$f(x;\mathbf{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$

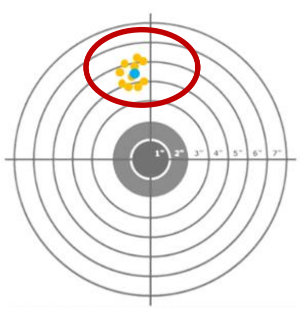
$f(x;\mathbf{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5$

Black curve: the true function f^*

Red curves: 5000 $\hat{f}$

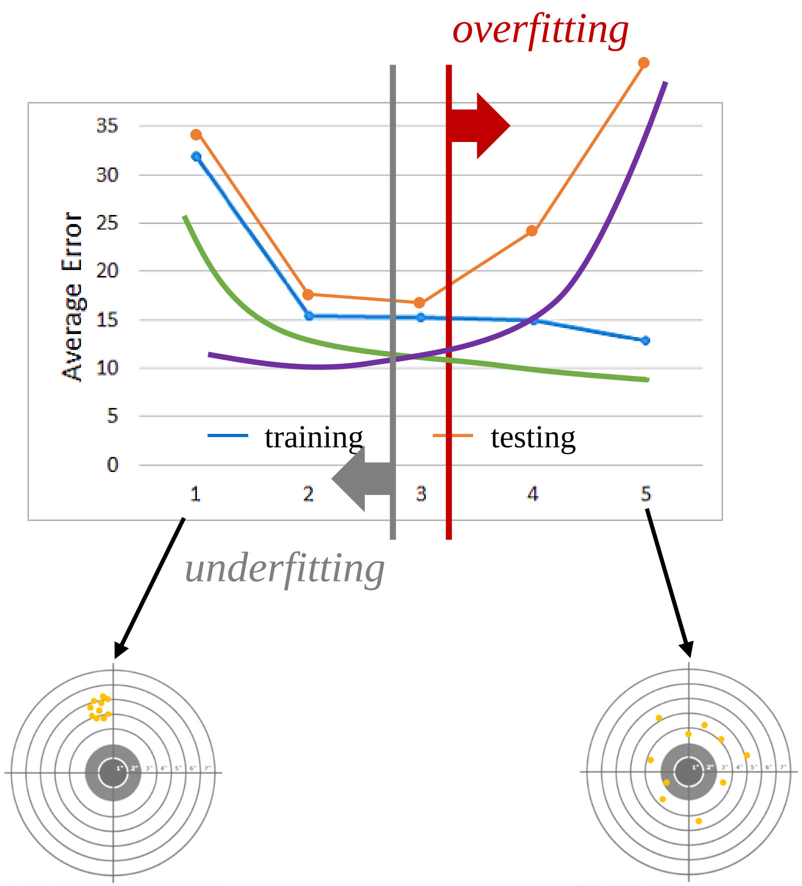
Blue curve: the average of 5000 $\hat{f}$

large bias



small bias

10.6.5 Diagnosis and what to do

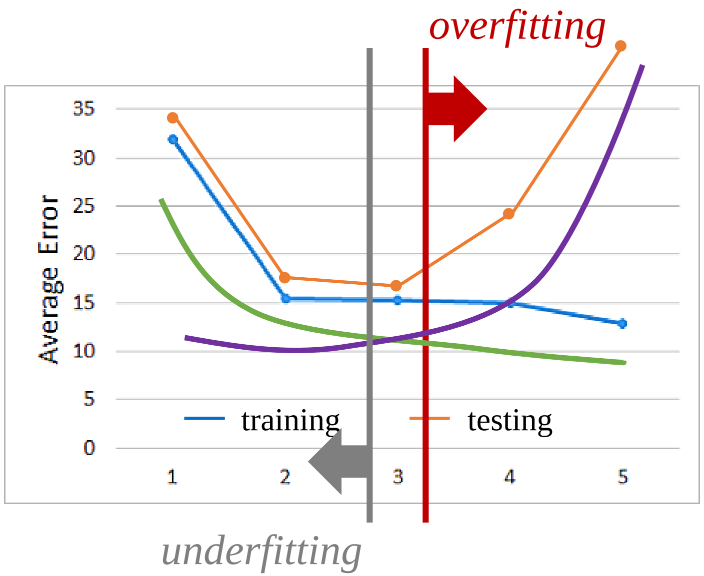


large bias
small variance

small bias
large variance

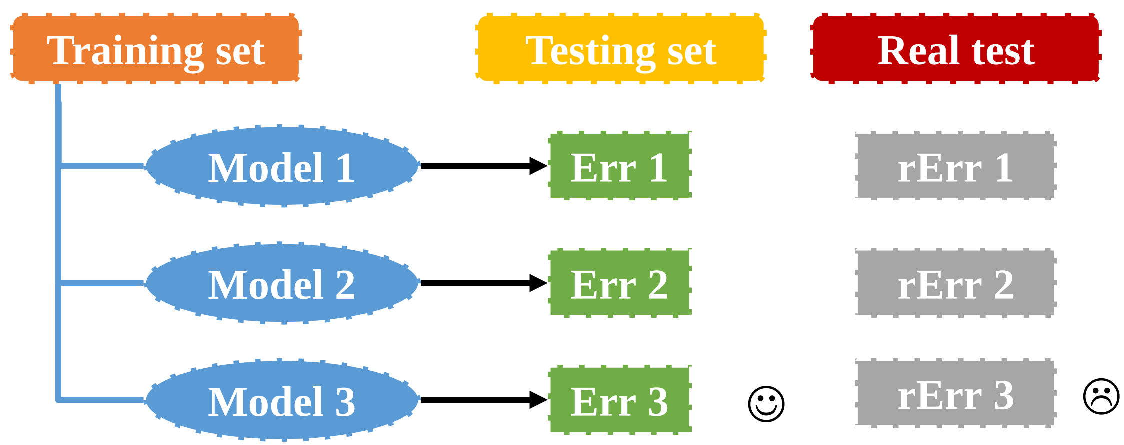
- Diagnosis:
  - If your model cannot even fit the training examples, then you have large bias ***(underfitting!)***
  - If you van fit the training data, but large error on testing data, then you probably have large variance ***(overfitting!)***
- What to do:
  - Redesign your model: add more features as input; a more complex model… ***(underfitting!)***
  - Collect more data; generate “fake” data; add regularizations (constraints on model space) ***(overfitting!)***

10.6.6 Model selection



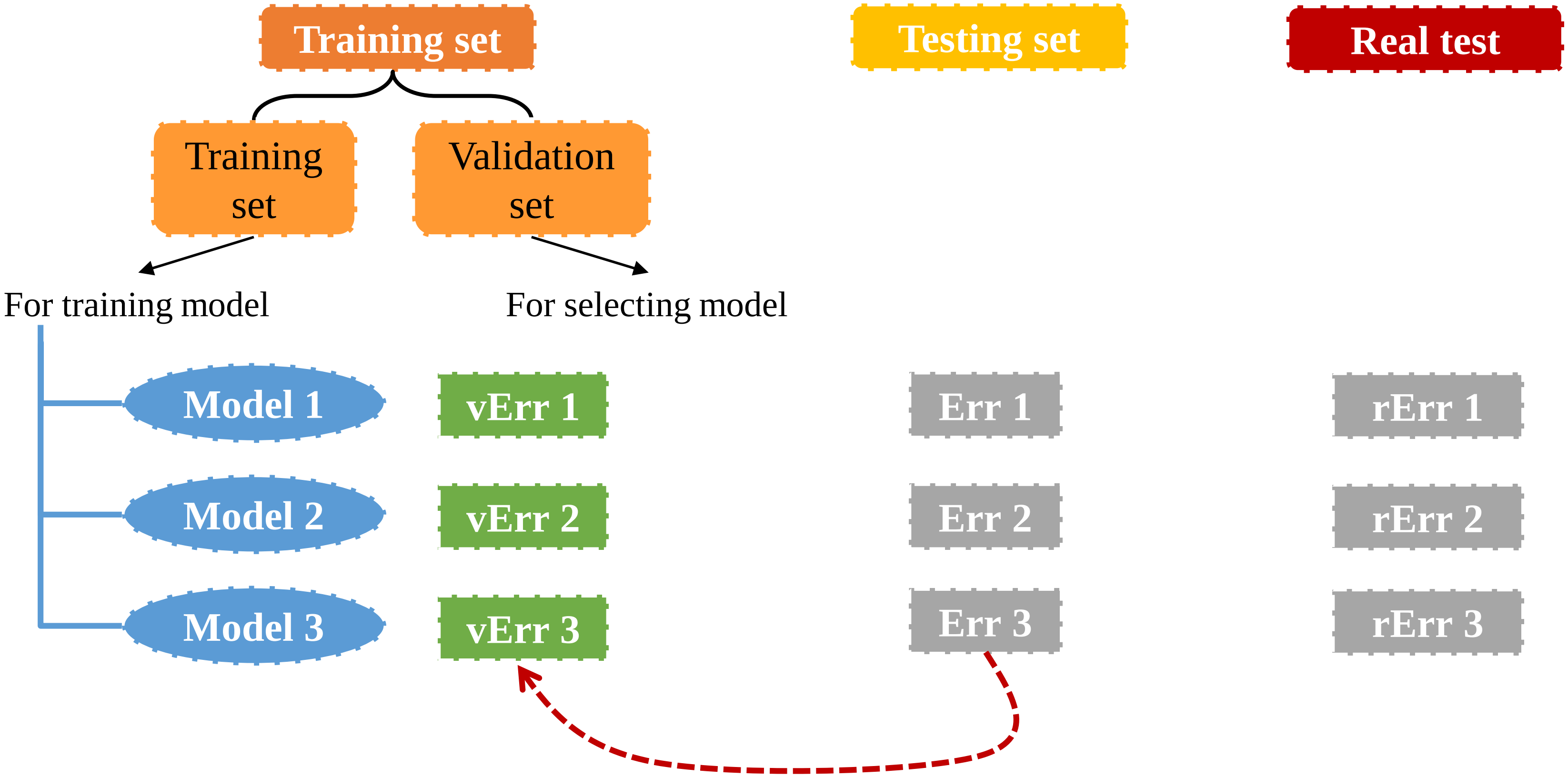
- There is usually a trade-off between bias and variance
- Select a model that balances two kinds of error to minimize total error

- What you should not do:



10.6.7 Cross validation

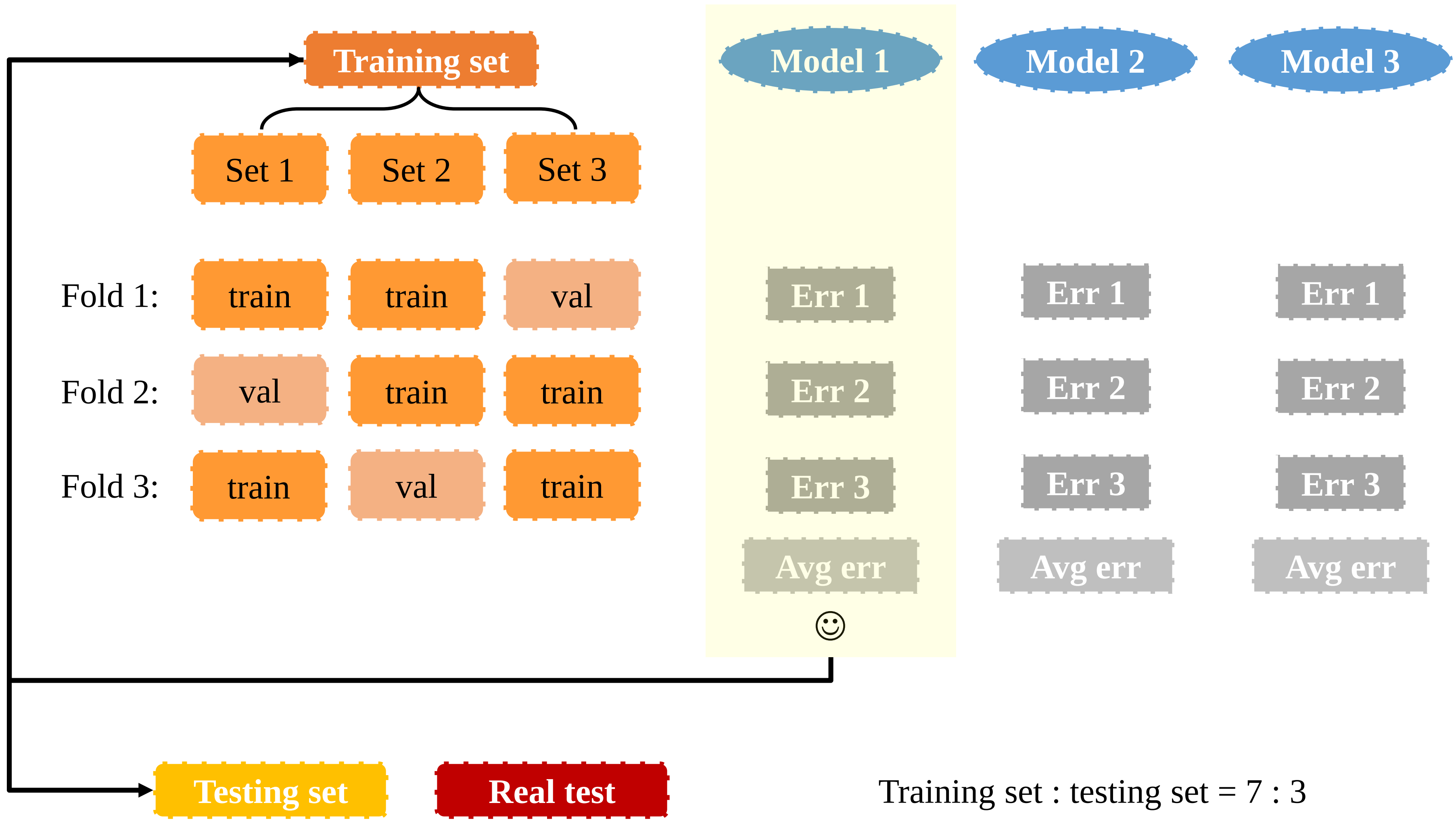
- What you should do:



not recommend:
using the results of testing set
to tune your model

10.6.8 K-fold Cross validation

- If we do three-fold cross validation:



Training set : testing set = 7 : 3

10.7 Performance measurements

input data	output variable	ground truth
$\{x_i\}_{i=1,\cdots,n}$	$\{z_i\}_{i=1,\cdots,n}$	$\{y_i\}_{i=1,\cdots,n}$
mean	$\bar{z}$	$\bar{y}$
std.	σ_z	σ_y

- Mean absolute error

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |z_i - y_i|$$

- Mean absolute percentage error

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{z_i - y_i}{y_i}\right| \cdot 100\%$$

  - MAPE is scale independent
  - It derives infinity or undefined value when $y_i = 0$
  - But for this study, it doesn't matter

- Coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(z_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  - From 0 to 1, higher is better
  - Defines the proportion that a model can explain of total variation

- Anomaly correlation coefficient

$$ACC = \frac{1}{n}\frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sigma_z \sigma_y}$$

  - From -1 to 1, higher is better
  - It indicates linear correlativity between two variables

- (Root) mean square error

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (z_i - y_i)^2$$

10.7 Performance measurements

	input data	output variable	ground truth
	$\{x_i\}_{i=1,\cdots,n}$	$\{z_i\}_{i=1,\cdots,n}$	$\{y_i\}_{i=1,\cdots,n}$
mean	$\bar{z}$		$\bar{y}$
std.	σ_z		σ_y

- (Root) mean square error

$$\boldsymbol{MSE} = \frac{1}{n}\sum_{i=1}^{n}(z_i - y_i)^2$$

$$\boldsymbol{MSE} = \frac{1}{n}\sum_{i=1}^{n}(z_i - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n}[(z_i - \bar{z}) - (y_i - \bar{y}) + (\bar{z} - \bar{y})]^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 + (\bar{z} - \bar{y}) + \cancel{\frac{2}{n}\sum_{i=1}^{n}[(z_i - \bar{z}) - (y_i - \bar{y})](\bar{z} - \bar{y})} - \frac{2}{n}\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})$$

$$\boldsymbol{MSE} = \sigma_z^2 + \sigma_y^2 + (\bar{z} - \bar{y}) - 2\sigma_z\sigma_y \boldsymbol{ACC} = \boldsymbol{E_m^2} + \boldsymbol{E_p^2}$$

where

$\boldsymbol{E_m^2} = (\bar{z} - \bar{y})$ indicates MSE by *mean difference*

$\boldsymbol{E_p^2} = \sigma_z^2 + \sigma_y^2 - 2\sigma_z\sigma_y \boldsymbol{ACC}$ denotes MSE caused by *pattern variation*