Video 26

Sunday, November 22, 2020 4:00 PM

Loss: squared error loss

Input: x

Prediction: y

Targets: conditional distribution $p_D(t|x)$

Minimizing the expected loss:

$$\mathbb{E}_{p_D}\left[(y-t)^2 \mid x\right] = \mathbb{E}\left[y^2 - 2yt + t^2 \mid x\right]$$

$(y-t)^2$: squared error loss

$$= y^2 - 2y\,\mathbb{E}[t|x] + \mathbb{E}[t^2|x]$$

$$= y^2 - 2y\,\mathbb{E}[t|x] + \mathbb{E}[t|x]^2 + \mathrm{Var}[t|x]$$

$\mathrm{Var}[z] = \mathbb{E}[z^2] - \mathbb{E}[z]^2$

$$= \left(y - \mathbb{E}[t|x]\right)^2 + \mathrm{Var}[t|x]$$

$\mathrm{Var}[t|x]$: Bayes error

y^*: best possible prediction

- $(y-y^*)^2 \geq 0$
- $\mathrm{Var}[t|x]$ does not depend on y

Bayes error: best possible generalization error we can achieve (if we model data perfectly)

$\approx$ best risk we can achieve

Treat y as a random variable

Experiment:

1. Sample a training set from p_D
2. Train a model (e.g., neural network)
3. Compute predictions on x

(for simplicity, omit the dependence on x)

$$\mathbb{E}\left((y-t)^2\right) = \mathbb{E}\left[(y-y^*)^2\right] + \mathrm{Var}(t)$$

$$= \mathbb{E}\left[y^2 - 2yy^* + y^{*2}\right] + \mathrm{Var}(t)$$

$$= \mathbb{E}[y^2] - 2y^*\,\mathbb{E}[y] + y^{*2} + \mathrm{Var}(t)$$

$$= y^{*2} + \mathbb{E}[y]^2 + \mathrm{Var}(y) - 2y^*\,\mathbb{E}[y] + \mathrm{Var}(t)$$

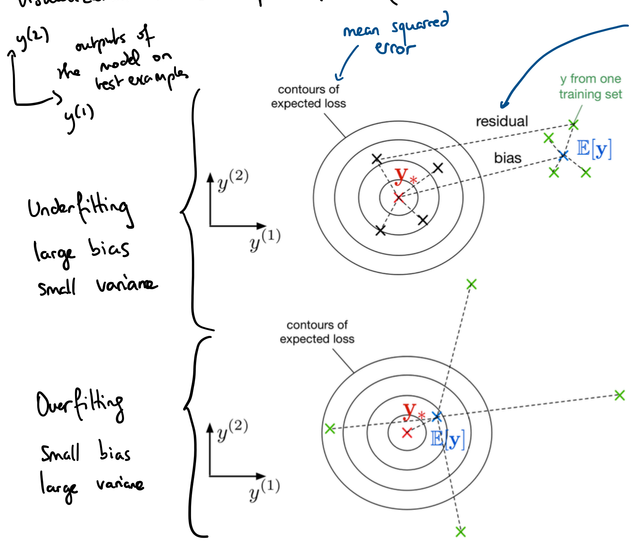
$$= y^{*2} - 2y^*\,\mathbb{E}[y] + \mathbb{E}[y]^2 + \mathrm{Var}(y) + \mathrm{Var}(t)$$

Bias-variance decomposition

$$\mathbb{E}\left[(y-t)^2\right] = \left(y^* - \mathbb{E}[y]\right)^2 + \mathrm{Var}(y) + \mathrm{Var}(t)$$

- $\left(y^* - \mathbb{E}[y]\right)^2$: bias — how far our model's average prediction
- $\mathrm{Var}(y)$: variance — of predictions explained by the choice of training set (how much does the model overfit to the train set)
- $\mathrm{Var}(t)$: Bayes error

Visualization in the output space (2 test examples)



Bias term $\|\mathbb{E}[y] - y_*\|^2$

Variance: spread of the green x

Bayes error: spread of the black x