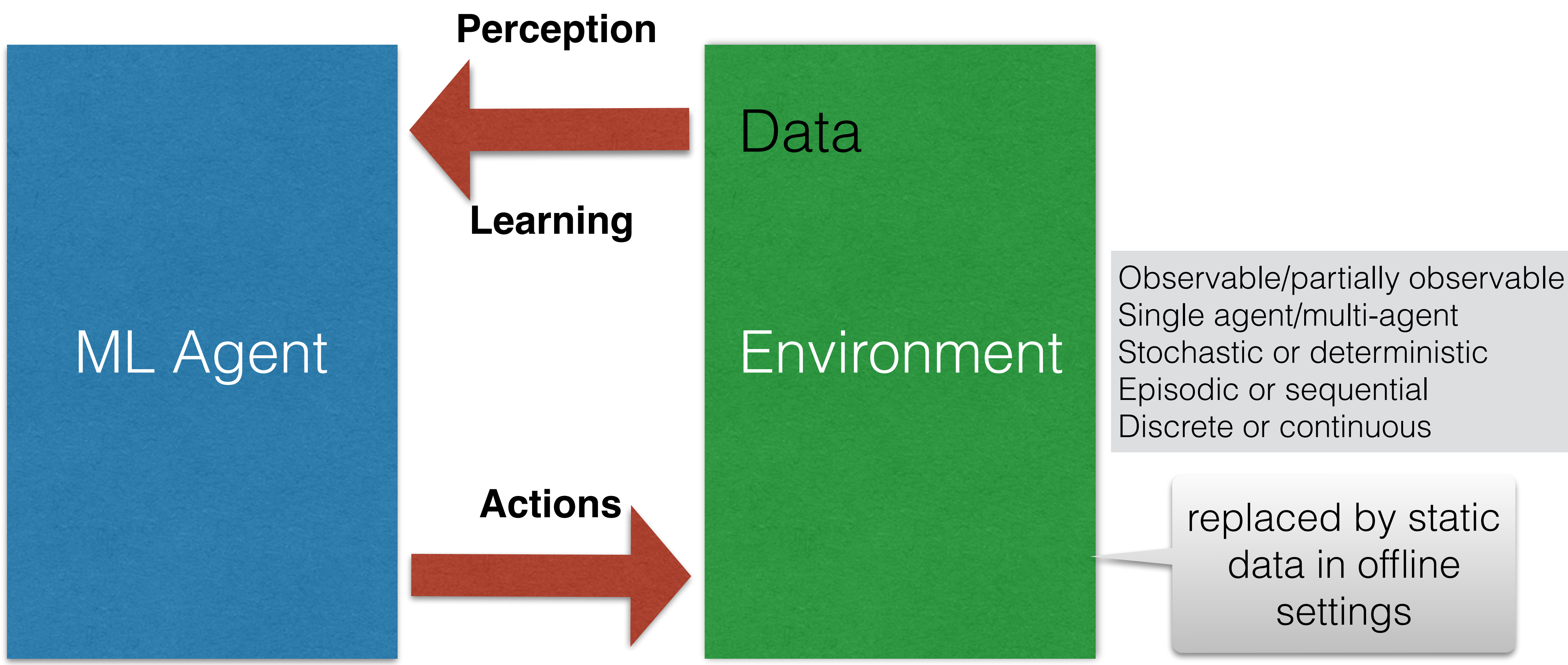
Guided Tour of Machine Learning in Finance

The bias-variance tradeoff

Igor Halperin

NYU Tandon School of Engineering, 2017

Generalization as a goal of learning



Learning = ability to generalize

= perform well on unseen data

Generalization error in regression

- Consider a **regression problem**

$$y = f(\boldsymbol{x}) + \varepsilon$$

- Here $\mathbb{E}[\varepsilon] = 0, \mathbb{E}\left[\varepsilon^2\right] = \sigma^2$

Generalization error in regression

- Consider a **regression problem**

$$y = f(\boldsymbol{x}) + \varepsilon$$

- Here $\mathbb{E}[\varepsilon] = 0, \mathbb{E}\left[\varepsilon^2\right] = \sigma^2$

- This produces $\mathbb{E}[y] = f(x), Var(y) = \sigma^2$

Generalization error in regression

- Consider a regression problem

$$y = f(\boldsymbol{x}) + \varepsilon$$

- Here $\mathbb{E}[\varepsilon] = 0, \mathbb{E}\left[\varepsilon^2\right] = \sigma^2$

- This produces $\mathbb{E}[y] = f(x), Var(y) = \sigma^2$

- We look for approximation $f(\boldsymbol{x}) \approx \hat{f}(\boldsymbol{x})$ such that the **generalization error** (expected loss) $\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right]$ is minimized over <u>all possible data</u>, both seen and unseen!

Generalization error in regression

- Consider a regression problem

$$y = f(\boldsymbol{x}) + \varepsilon$$

- Here $\mathbb{E}[\varepsilon] = 0, \mathbb{E}\left[\varepsilon^2\right] = \sigma^2$

- This produces $\mathbb{E}[y] = f(x), Var(y) = \sigma^2$

- We look for approximation $f(\boldsymbol{x}) \approx \hat{f}(\boldsymbol{x})$ such that the **generalization error** (expected loss) $\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right]$ is minimized over <u>all possible data</u>, both seen and unseen!

- Evaluate: $\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathbb{E}\left[y^2\right] + \mathbb{E}\left[\hat{f}^2\right] - 2\mathbb{E}\left[y\hat{f}\right]$

Generalization error in regression

- Consider a regression problem

$$y = f(\boldsymbol{x}) + \varepsilon$$

- Here $\mathbb{E}[\varepsilon] = 0, \mathbb{E}\left[\varepsilon^2\right] = \sigma^2$

- This produces $\mathbb{E}[y] = f(x), Var(y) = \sigma^2$

- We look for approximation $f(\boldsymbol{x}) \approx \hat{f}(\boldsymbol{x})$ such that the **generalization error** (expected loss) $\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right]$ is minimized over <u>all possible data</u>, both seen and unseen!

- Evaluate: $\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathbb{E}\left[y^2\right] + \mathbb{E}\left[\hat{f}^2\right] - 2\mathbb{E}\left[y\hat{f}\right] =$

$$Var(y) + \mathbb{E}[y]^2 + Var(\hat{f}) + \mathbb{E}\left[\hat{f}\right]^2 - 2\mathbb{E}\left[y\hat{f}\right] =$$

$y = f(\boldsymbol{x}) + \varepsilon$

Generalization error in regression

- Consider a regression problem

$$y = f(\boldsymbol{x}) + \varepsilon$$

- Here $\mathbb{E}[\varepsilon] = 0, \mathbb{E}\left[\varepsilon^2\right] = \sigma^2$

- This produces $\mathbb{E}[y] = f(x), Var(y) = \sigma^2$

- We look for approximation $f(\boldsymbol{x}) \approx \hat{f}(\boldsymbol{x})$ such that the **generalization error** (expected loss) $\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right]$ is minimized over <u>all possible data</u>, both seen and unseen!

- Evaluate: $\mathbb{E}\left[\left(y - \hat{f}(x)\right)^2\right] = \mathbb{E}\left[y^2\right] + \mathbb{E}\left[\hat{f}^2\right] - 2\mathbb{E}\left[y\hat{f}\right] =$

$$Var(y) + \mathbb{E}[y]^2 + Var(\hat{f}) + \mathbb{E}\left[\hat{f}\right]^2 - 2\mathbb{E}\left[y\hat{f}\right] =$$

$$Var(y) + Var(\hat{f}) + \mathbb{E}\left[f - \hat{f}\right]^2$$

$y = f(\boldsymbol{x}) + \varepsilon$

The Bias-Variance Decomposition

This is is bias-variance decomposition:

$$\boxed{\mathbb{E}\left[\left(y - \hat{f}(\boldsymbol{x})\right)^2\right] = (bias)^2 + variance + noise}$$

Here:

$$(bias)^2 = \mathbb{E}\left[f - \hat{f}\right]^2$$

$$variance = Var(\hat{f})$$

$$\text{noise} = Var(y) = \sigma^2$$

The Bias-Variance Decomposition

This is is bias-variance decomposition:

$$\boxed{\mathbb{E}\left[\left(y-\hat{f}(\boldsymbol{x})\right)^2\right]=(bias)^2+variance+noise}$$

Here:

$$(bias)^2=\mathbb{E}\left[f-\hat{f}\right]^2$$

$$variance=Var(\hat{f})$$

$$\text{noise}=Var(y)=\sigma^2$$

Bias: the (square of) expected difference of approximate predictor $\hat{f}(\boldsymbol{x})$ from the “true” predictor $f(\boldsymbol{x})$

The Bias-Variance Decomposition

This is is bias-variance decomposition:

$$\boxed{\mathbb{E}\left[\left(y-\hat{f}(\boldsymbol{x})\right)^2\right]=(bias)^2+variance+noise}$$

Here:

$$(bias)^2=\mathbb{E}\left[f-\hat{f}\right]^2$$

$$variance=Var(\hat{f})$$

$$\text{noise}=Var(y)=\sigma^2$$

Bias: the (square of) expected difference of approximate predictor $\hat{f}(\boldsymbol{x})$ from the "true" predictor $f(\boldsymbol{x})$

Variance: sensitivity of $\hat{f}(\boldsymbol{x})$ to the choice of data set

The Bias-Variance Decomposition

This is is bias-variance decomposition:

$$\mathbb{E}\left[\left(y-\hat{f}(\boldsymbol{x})\right)^2\right]=(bias)^2+variance+noise$$

Here:

$$(bias)^2=\mathbb{E}\left[f-\hat{f}\right]^2$$

$$variance=Var(\hat{f})$$

$$\mathrm{noise}=Var(y)=\sigma^2$$

Bias: the (square of) expected difference of approximate predictor $\hat{f}(\boldsymbol{x})$ from the "true" predictor $f(\boldsymbol{x})$

Variance: sensitivity of $\hat{f}(\boldsymbol{x})$ to the choice of data set

Noise: a property of data; beyond our control

Control question

Q: Select all correct statements

1. $variance = \mathbb{E}\left[f - \hat{f} \right]^2$ is a measure of variability of a model.

2. $variance = Var(\hat{f})$ determines sensitivity of a model $\hat{f}(\boldsymbol{x})$ with respect to the choice of a dataset.

3. $\left(bias \right)^2 = \mathbb{E}\left[f - \hat{f} \right]^2$ is a measure of flexibility of a model $\hat{f}(\boldsymbol{x})$ in matching typical data that can be observed for a given (and unknown) data-generating mechanism that corresponds to a "true" (and unknown) data model $f(\boldsymbol{x})$

Correct answers: 2, 3.

The Bias-Variance Tradeoff

$$\mathbb{E}\left[\left(y-\hat{f}(\boldsymbol{x})\right)^2\right]=\left(bias\right)^2+variance+noise$$

- This relation is mostly of a theoretical value: we do not know how to compute the bias and variance!
- But it show a general decomposition of expected loss of a generic regression model.

The Bias-Variance Tradeoff

$$\mathbb{E}\left[\left(y-\hat{f}(\boldsymbol{x})\right)^2\right]=(bias)^2+variance+noise$$

- This relation is mostly of a theoretical value: we do not know how to compute the bias and variance!
- But it show a general decomposition of expected loss of a generic regression model.
- **Bias-variance tradeoff:**
  - <u>Complex models</u> (more features) tend to have <u>low bias</u> and <u>high variance</u>

The Bias-Variance Tradeoff

$$\mathbb{E}\left[\left(y-\hat{f}(\boldsymbol{x})\right)^2\right]=(bias)^2+variance+noise$$

- This relation is mostly of a theoretical value: we do not know how to compute the bias and variance!
- But it show a general decomposition of expected loss of a generic regression model.
- **Bias-variance tradeoff:**
  - <u>Complex models</u> (more features) tend to have <u>low bias</u> and <u>high variance</u>
  - <u>Simple models</u> (less features) tend to have <u>high bias</u> and <u>low variance</u>

The Bias-Variance Tradeoff

$$\boxed{\mathbb{E}\left[\left(y-\hat{f}(\boldsymbol{x})\right)^2\right]=(bias)^2+variance+noise}$$

- This relation is mostly of a theoretical value: we do not know how to compute the bias and variance!
- But it show a general decomposition of expected loss of a generic regression model.
- **Bias-variance tradeoff:**
  - <u>Complex models</u> (more features) tend to have <u>low bias</u> and <u>high variance</u>
  - <u>Simple models</u> (less features) tend to have <u>high bias</u> and <u>low variance</u>
- Important to control **model complexity** to have an optimal tradeoff!

Control question

Q: Assume you have a complex model with a low bias and a high variance, and you want to reduce the variance. What do you think would be the right way to proceed?

A1: Use a simple model instead. Who needs complex models in Finance?

A2. Add more predictors.

A3. Try to somehow bound the values of model parameters, so that the model outputs would vary less with a variation of the input data.

Correct answer: A3.