Outline

Generalization error

Generalization error

- How accurately an algorithm is able to predict outcomes for previously unseen data

Generalization error

- How accurately an algorithm is able to predict outcomes for previously unseen data
  - $\rightarrow$ Risk or out-of-sample error

Generalization error

- How accurately an algorithm is able to predict outcomes for previously unseen data
  - $\rightarrow$ Risk or out-of-sample error
- Main causes:
  1) **Model complexity** $\rightarrow$ Reproduce the underlying distribution
  2) **Data availability** $\rightarrow$ Representative of the distribution to model

Generalization error

- How accurately an algorithm is able to predict outcomes for previously unseen data
  - → Risk or out-of-sample error
- Main causes:
  1) **Model complexity** → Reproduce the underlying distribution
  2) **Data availability** → Representative of the distribution to model
- Sources of error:
  1) **Bias** → Systematic assumptions by the model (*underfitting*)
  2) **Variance** → Sensitivity to training data (*overfitting*)

Bias-variance decomposition

Bias-variance decomposition

Regression task:

Bias-variance decomposition

Regression task:

- There exists a true (but unknown) function $f(\mathbf{x})$

Bias-variance decomposition

Regression task:

- There exists a true (but unknown) function $f(\mathbf{x})$
- **Premise**: approximate $\hat{f}(\mathbf{x})$ with set $\mathcal{D}$ generated by $f(\mathbf{x})$

Bias-variance decomposition

Regression task:

- There exists a true (but unknown) function $f(\mathbf{x})$
- **Premise**: approximate $\hat{f}(\mathbf{x})$ with set $\mathcal{D}$ generated by $f(\mathbf{x})$
  - $\rightarrow$ Random variations selecting $\mathcal{D} \Rightarrow$ better or worse approximation

Bias-variance decomposition

Regression task:

- There exists a true (but unknown) function $f(\mathbf{x})$
- **Premise**: approximate $\hat{f}(\mathbf{x})$ with set $\mathcal{D}$ generated by $f(\mathbf{x})$
  - $\rightarrow$ Random variations selecting $\mathcal{D} \Rightarrow$ better or worse approximation
- **Effectiveness** $\Rightarrow$ Deviation from desired optimal averaged over all $\mathcal{D}$ sets

Bias-variance decomposition

Regression task:

- There exists a true (but unknown) function $f(\mathbf{x})$
- **Premise**: approximate $\hat{f}(\mathbf{x})$ with set $\mathcal{D}$ generated by $f(\mathbf{x})$
  - $\rightarrow$ Random variations selecting $\mathcal{D} \Rightarrow$ better or worse approximation
- **Effectiveness** $\Rightarrow$ Deviation from desired optimal averaged over all $\mathcal{D}$ sets

$$\mathrm{Err}(\mathbf{x}) = \underbrace{\left(\mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x};\mathcal{D})] - f(\mathbf{x})\right)^2}_{\textit{Bias}^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\hat{f}(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x};\mathcal{D})]\right]}_{\textit{Variance}} + \underbrace{\sigma_e^2}_{\textit{Irreducible noise}}$$

Bias-variance decomposition

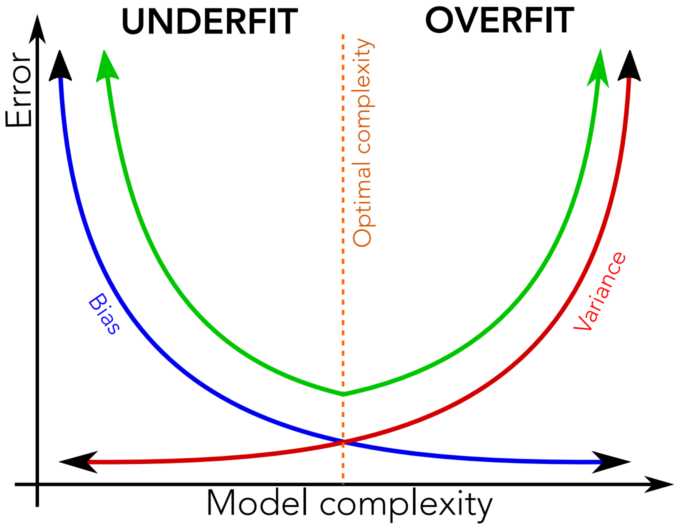
Regression task:

- There exists a true (but unknown) function $f(\mathbf{x})$
- **Premise**: approximate $\hat{f}(\mathbf{x})$ with set $\mathcal{D}$ generated by $f(\mathbf{x})$
  - $\rightarrow$ Random variations selecting $\mathcal{D}$ $\Rightarrow$ better or worse approximation
- **Effectiveness** $\Rightarrow$ Deviation from desired optimal averaged over all $\mathcal{D}$ sets

$$\mathrm{Err}(\mathbf{x}) = \underbrace{\left(\mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x};\mathcal{D})] - f(\mathbf{x})\right)^2}_{Bias^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\hat{f}(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x};\mathcal{D})]\right]}_{Variance} + \underbrace{\sigma_e^2}_{Irreducible\ noise}$$

- **Bias**: Systematic deviation from $f(\mathbf{x})$
- **Variance**: Dependence on the data sampling in $\mathcal{D}$

Bias-variance tradeoff



Curse of dimensionality

Curse of dimensionality

- Amount of features d used $\Rightarrow \mathbf{x} \in \mathbb{R}^d$

Curse of dimensionality

- Amount of features d used $\Rightarrow \mathbf{x} \in \mathbb{R}^d$
- Theoretically, estimator benefits from the use of larger d

Curse of dimensionality

- Amount of features d used $\Rightarrow \mathbf{x} \in \mathbb{R}^d$
- Theoretically, estimator benefits from the use of larger d
  - $\rightarrow$ In practise, may become an issue (high-dimensional spaces)

Curse of dimensionality

- Amount of features d used $\Rightarrow \mathbf{x} \in \mathbb{R}^d$
- Theoretically, estimator benefits from the use of larger d
  - $\rightarrow$ In practise, may become an issue (high-dimensional spaces)
  - $\rightarrow$ Data increasingly becomes sparse

Curse of dimensionality

- Amount of features d used $\Rightarrow \mathbf{x} \in \mathbb{R}^d$
- Theoretically, estimator benefits from the use of larger d
  - $\rightarrow$ In practise, may become an issue (high-dimensional spaces)
  - $\rightarrow$ Data increasingly becomes sparse
  - $\rightarrow$ Severe overfitting risk

Curse of dimensionality

- Amount of features d used $\Rightarrow \mathbf{x} \in \mathbb{R}^d$
- Theoretically, estimator benefits from the use of larger d
  - $\rightarrow$ In practise, may become an issue (high-dimensional spaces)
  - $\rightarrow$ Data increasingly becomes sparse
  - $\rightarrow$ Severe overfitting risk
- Possible strategies to palliate it:

Curse of dimensionality

- Amount of features d used $\Rightarrow \mathbf{x} \in \mathbb{R}^d$
- Theoretically, estimator benefits from the use of larger d
  - $\rightarrow$ In practise, may become an issue (high-dimensional spaces)
  - $\rightarrow$ Data increasingly becomes sparse
  - $\rightarrow$ Severe overfitting risk
- Possible strategies to palliate it:
  - $\rightarrow$ **Adequate design** of the feature extraction stage

Curse of dimensionality

- Amount of features d used $\Rightarrow \mathbf{x} \in \mathbb{R}^d$
- Theoretically, estimator benefits from the use of larger d
  - $\rightarrow$ In practise, may become an issue (high-dimensional spaces)
  - $\rightarrow$ Data increasingly becomes sparse
  - $\rightarrow$ Severe overfitting risk
- Possible strategies to palliate it:
  - $\rightarrow$ **Adequate design** of the feature extraction stage
  - $\rightarrow$ **Feature selection**: Select subset of features $d' < d$

Curse of dimensionality

- Amount of features d used $\Rightarrow \mathbf{x} \in \mathbb{R}^d$
- Theoretically, estimator benefits from the use of larger d
  - $\rightarrow$ In practise, may become an issue (high-dimensional spaces)
  - $\rightarrow$ Data increasingly becomes sparse
  - $\rightarrow$ Severe overfitting risk
- Possible strategies to palliate it:
  - $\rightarrow$ **Adequate design** of the feature extraction stage
  - $\rightarrow$ **Feature selection**: Select subset of features $d' < d$
  - $\rightarrow$ **Model regularization**: Add constraints/penalties to the loss function

T2: Computational learning

Fundamentos del Aprendizaje Automático

Curso 2025/2026