CIE 555
Neural Networks and Deep Learning

Machine Learning Basics

1

Overview

- Review of some Linear Algebra basics (selected sections – see next slide)
- Review of the basics of Probability and Information Theory (selected sections – see next slide)
- Capacity, Overfitting and Underfitting (section 5.2)
- Hyperparameters and Validation Sets (section 5.3)
- Estimators, Bias and Variance (section 5.4)
- Consistency (section 5.4.5)

2

Review

- Review of some Linear Algebra basics. In particular we discussed
  - Scalars, Vectors, Matrices and Tensors (section 2.1)
  - Norms (section 2.5)
  - Other concepts will be introduced in upcoming lectures where needed
- Review of the basics of Probability and Information Theory
  - Random Variables (section 3.2)
  - Probability Distributions (section 3.3)
  - Marginal Probability (section 3.4)
  - Conditional Probability (section 3.5)
  - The Chain Rule of Conditional Probabilities (section 3.6)
  - Independence and Conditional Independence (section 3.7)
  - Expectation and Variance (section 3.8)
  - Bayes' Rule (section 3.11)

3

Machine Learning

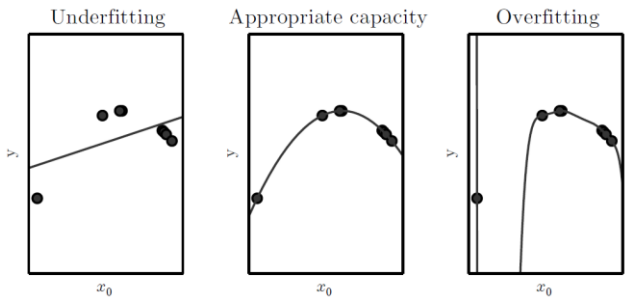
- "A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P* , if its performance at tasks in *T* , as measured by *P* , improves with experience *E*." Mitchell (1997)
- **Tasks:** described in terms of how the machine learning system should process an **example** (a collection of features that have been quantitatively measured from some object or event).
- Example tasks: classification, regression, transcription, machine translation, anomaly detection, etc.
- **Performance:** specific to the *task T* (e.g. accuracy, error rate)
- **The Experience, *E:*** Most machine learning algorithms simply experience a dataset.

Goodfellow, Bengio, Courville 2016

4

Capacity, Overfitting and Underfitting

- **Underfitting:** occurs when the model is not able to obtain a sufficiently low error value on the training set.
- **Overfitting:** occurs when the gap between the training error and test error is too large.
- We can control whether a model is more likely to overfit or underfit by altering its **capacity** (Informally, a model's capacity is its ability to fit a wide variety of functions.).

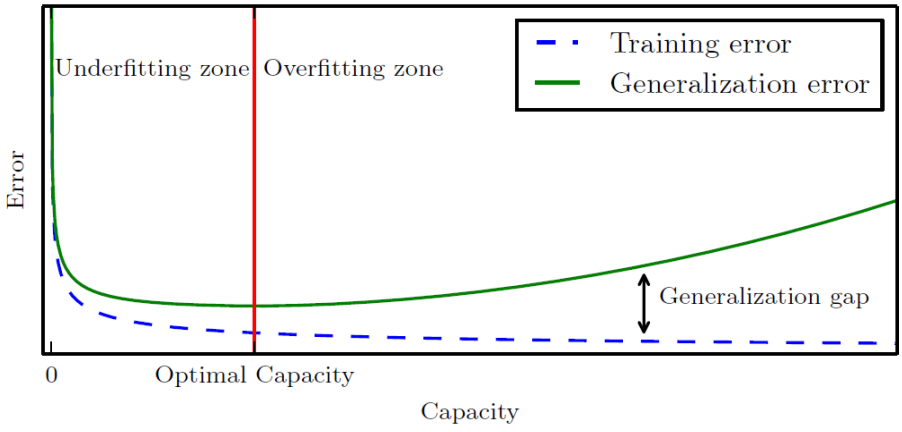


Underfitting and Overfitting in Polynomial Estimation

Goodfellow, Bengio, Courville 2016

5

Capacity, Overfitting and Underfitting



Generalization and Capacity

Goodfellow, Bengio, Courville 2016

6

Non-parametric models

- Parametric models learn a function described by a parameter vector whose size is ***finite and fixed*** before any data is observed.
- Non-parametric models have no such limitation.
- However, we can also design practical non-parametric models by making their complexity a function of the training set size (example: nearest neighbor regression).
- When asked to classify a test point x , the model looks up the nearest entry in the training set and returns the associated regression target.
- While simpler functions are more likely to generalize (to have a small gap between training and test error) we must still choose a sufficiently complex hypothesis to achieve low training error.

7

Hyperparameters and Validation Sets

- **Hyperparameters**: settings that designers use to control the behavior of the learning algorithm.
  - In polynomial regression: the degree of the polynomial is a hyperparameter (capacity hyperparmeter)
  - The learning rate λ
- Typically, it is inappropriate to learn hyperparameters on the training set. Why?
  - We need a **validation set**
  - no example from the test set can be used in the validation set. (why?)
  - we always construct the validation set from the *training data* (split training dataset into two disjoint subsets).
  - Typically, one uses about 80% of the training data for training and 20% for validation.
  - the validation set error will *underestimate* the generalization error, though typically by a smaller amount than the training error.

Goodfellow, Bengio, Courville 2016

8

Cross-Validation

- Dividing the dataset into a fixed training set and a fixed test set can be problematic if it results in the test set being small.
- A small test set implies *statistical uncertainty* around the *estimated average validation error*, making it difficult to claim that algorithm *A* works better than algorithm *B* on the given task.
- When the dataset is too small, are alternative procedures enable one to *use all of the examples in the estimation of the mean test error*, at the price of increased computational cost.
- These procedures are based on the idea of *repeating the training and testing computation on different randomly chosen subsets or splits of the original dataset*.

Goodfellow, Bengio, Courville 2016

9

Cross-Validation

- The most common of these is *the k-fold cross-validation procedure*, in which a partition of the dataset is formed by splitting it into k non-overlapping subsets.
- The test error may then be estimated by taking the average test error across *k* trials.
- One problem is that there exist no unbiased estimators of the variance of such average error estimators (Bengio and Grandvalet, 2004), but approximations are typically used.

Goodfellow, Bengio, Courville 2016

10

Estimators, Bias and Variance

- Estimation
  - E.g. estimating the weights in the linear regression example
  - Let $\{x^{(1)}, \ldots, x^{(m)}\}$ be a set of *m* independent and identically distributed data points
  - A point estimator is any function $\hat{\theta}_m = g(x^{(1)}, \ldots, x^{(m)})$
  - a good estimator is a function whose output is close to the true underlying function θ that generated the training data.

Goodfellow, Bengio, Courville 2016

11

Bias

- Bias
  - measures the expected deviation from the true value of the function or parameter.
  - Defined as: $bias(\hat{\theta}_m) = \mathbb{E}(\hat{\theta}_m) - \theta$ where $\mathbb{E}(\hat{\theta}_m)$: expectation over the data
  - An estimator $\hat{\theta}_m$ is *unbiased* when $bias(\hat{\theta}_m) = 0$
  - An estimator $\hat{\theta}_m$ is said to be **asymptotically unbiased** if $\lim_{m\to\infty} bias(\hat{\theta}_m) = 0$ $(i.e. \lim_{m\to\infty} \mathbb{E}(\hat{\theta}_m) = \theta)$
  - *Asymptotically unbiased estimators:* estimators whose bias goes to 0 as the sample size goes to infinity.

Goodfellow, Bengio, Courville 2016

12

Variance and Standard Error

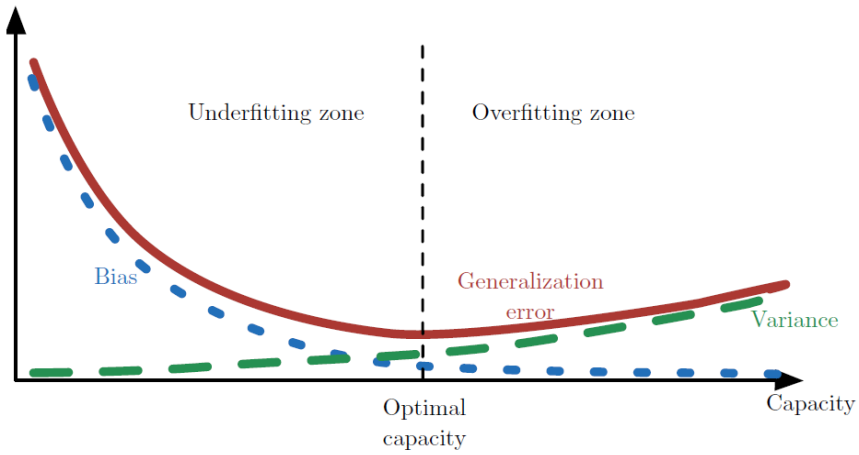
- Measures how we would expect the estimate we compute from data to vary as we independently resample the dataset from the underlying data generating process.
- Simply the variance $Var(\hat{\theta})$
- *Standard error:* the square root of the variance.

Goodfellow, Bengio, Courville 2016

13

Trading off Bias and Variance

- tightly linked to the concepts of capacity, underfitting and overfitting
- increasing capacity tends to increase variance and decrease bias



Goodfellow, Bengio, Courville 2016

14

Consistency

- *Consistency* ensures that the bias and variance induced by the estimator diminishes as the number of data examples grows.
- We would like that $\operatorname{plim}_{m \to \infty} \hat{\theta}_m = \theta$ (plim: indicates convergence in probability)
- meaning that for any $\in > 0$, $P(\left|\hat{\theta}_m - \theta\right| > \in) \rightarrow 0$ as m $\rightarrow \infty$
- The estimator will be consistent if it is asymptotically unbiased, ***and*** its variance $\rightarrow 0$ as m $\rightarrow \infty$.

Goodfellow, Bengio, Courville 2016

15