Machine Learning

February 14, 2016

1 Basic outline of the ML lecture in WS15/16

The first part of the lecture followed the book *Pattern Recognition and Machine Learning* (PRML) by C. M. Bishop.

1.1 Lecture I

Introduction and first example of machine learning.

- Motivation:
  - What is “Machine Learning”?
    Learning according to Wikipedia:

    > Learning is the act of acquiring new, or modifying and reinforcing, existing knowledge, behaviors, skills, values, or preferences and may involve synthesizing different types of information

    Machine Learning now tries to equip machines with the ability to learn. In 1959 Arthur Samuel wrote the first self-learning program that could play Checkers. Accordingly he defined machine learning as

    > Field of study that gives computers the ability to learn without being explicitly programmed.

    A more formal definition in operational terms is provided by Tom Mitchell:

    > 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.

  - Examples:
    - *Classification:* Spam filtering, Face recognition, Medical diagnosis . . .
    - *Regression:* Weather prediction, Economic forecasts . . .

Some example data sets can be found at:
`http://archive.ics.uci.edu/ml/`: Spambase Data Set, Breast Cancer Wisconsin (Diagnostic) Data Set, El Nino Data Set
`http://deeplearning.net/datasets/`: MNIST
Introduced basic terminology and notation (PRML ch. 1(.0), i.e. pages 1-5):
    * Training and test data, Generalization
    * Pre-processing and feature extraction
    * Supervised vs. unsupervised learning
    * Classification and regression
  - Example of polynomial regression illustrates overfitting and regularization (PRML ch. 1.1).

1.2 Lecture II

Basic terminology and probabilistic formulation of machine learning

- Probability and decision theory
  - Explained how machine learning can be understood as a statistical problem (PRML 1.2.5)
  - Introduced decision theory and the concept of *loss functions* (PRML 1.5.5, regression only)
- Curse of dimensionality (PRML 1.4)
- Model selection and bias variance decomposition (PRML 1.3 and 3.2)

  Note: A review of basic probability theory was given in one of the exercise sessions. It covered PRML ch. 1.2-1.2.3 and simple examples of conjugate priors from PRML ch. 2.1.

1.3 Lecture III

Short recap of cross-validation:

- Training error is overly optimistic
- Cross-validation involves a bias-variance tradeoff wrt. number of folds

Linear models for regression: Revisited the polynomial regression example and formally solved it.

- Maximum likelihood and least squares (PRML 3.1.1 and 3.1.2)
- Regularized linear regression (PRML 3.1.4)

  Bayesian approach to machine learning motivates choice of regularizer as choice of prior as it can be understood as the MAP (Maximum a-posteriori) solution.
- Bayesian linear regression (PRML 3.3.1 and 3.3.2)

1.4 Lecture IV

- Bayesian model comparison and Ockham's razor (PRML 3.4)

  Stressed role of model evidence (PRML page 161) and discussed figures PRML 3.12 and 3.13.

- Conjugate priors for the Gaussian distribution (PRML 2.3.6, only mean and precision alone, joint prior mentioned in exercises).

- Formal solution for Bayesian linear regression (PRML 3.3.1). Also need to know formulas for marginal and conditional distributions of multivariate Gaussians (PRML Box on page 89-90).

1.5 Lecture V

Wrap up of first part by short recall of covered topics:

- Machine learning as a statistical modeling

- Problem of overfitting and potential remedies:

  - Regularization (often combined with cross-validation)
  - Cross-Validation (general method which applies widely)
  - Bayesian model selection (Bayesian approach of treating parameters as random variables handles uncertainty in a principled way and automatically punishes overly complex models).

- Linear regression: Includes polynomial regression by choosing appropriate basis function. What makes it linear is that basis functions are not learned, i.e. not adjusted to the data.

Finished Bayesian linear regression:

- Predictive distribution (PRML 3.3.2)

- Computed the evidence (PRML 3.5.1)