Introduction to Statistical Machine Learning

Cheng Soon Ong & Christian Walder

Machine Learning Research Group
Data61 | CSIRO
and
College of Engineering and Computer Science
The Australian National University

Canberra
February – June 2017

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

Part XXII

Discussion and Summary

Flavour of this course

- Formalise intuitions about problems
- Use language of mathematics to express models
- Geometry, vectors, linear algebra for reasoning
- Probabilistic models to capture uncertainty
- Calculus to identify good parameters
- Design and analysis of algorithms
- Numerical algorithms in python
- Understand the choices when designing machine learning methods

Probability Theory

- Frequentist vs. Bayes approach
- Conditional Probability
- Bayes Theorem
- Discrete vs. continuous random variables
- Distributions (Gaussian, Bernoulli, Binomial, Beta, . . .)
- Multivariate Distributions
- Change of Variables
- Conjugate Priors

(Chapter 2)

Ong & Walder
Data61 | CSIRO
The Australian National University

Linear Algebra and Calculus

- Vector Space
- Matrix-Vector Multiplication $=$ Linear Combination
- Projection
- Positive (Semi)-definite Matrix
- Rank, Determinant, Trace, Inverse
- Eigenvectors, Eigenvalues
- Eigenvector Decomposition
- Singular Value Decomposition
- Directional Derivative
- Gradient Calculation

Optimisation

- Gradient descent and friends
- Linear programming (linear objective function, linear constraints)
- Quadratic programming (quadratic objective function, linear constraints)
- Nonlinear programming (nonlinear objective function, nonlinear constraints)
- Convex programming (objective function is convex, constraints, if any, form a convex set)
- Stochastic programming (some of the constraints or parameters depend on random variables)
- Dynamic programming (e.g. Hidden Markov Model)

Probabilistic Graphical Models

- Joint Probability factorises.
- Conditional Independence
- Independence Structure or "the absence of edges"
- Directed, Undirected and Factor Graphs
- Bayesian Network, Blocked Path and d-separation
- Markov Random Field, (maximal) Cliques
- Factor Graphs are Bipartite Graphs

(Chapter 8)

What is Machine Learning?

Definition (Mitchell, 1998)

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.

Learning

- Examples
- General Setup
- Inductive Bias
- Restricted Hypothesis Space
- Importance of understanding the restrictions and whether they are appropriate
- Do not train on the test set

(Chapter 1)

Strategy in this course

- Estimate best predictor = training = learning
  Given data $(x_1, y_1), \ldots, (x_n, y_n)$, find a predictor $f_{\mathbf{w}}(\cdot)$.
  1. Identify the type of input x and output y data
  2. Propose a mathematical model for $f_{\mathbf{w}}$
  3. Design an objective function or likelihood
  4. Calculate the optimal parameter ($\mathbf{w}$)
  5. Model uncertainty using the Bayesian approach
  6. Implement and compute (the algorithm in python)
  7. Interpret and diagnose results

Tasks

- Regression
- Classification: binary and multiclass
- Dimensionality reduction
- Clustering
- Structured prediction

Models for Decision Problems

Given the input space $\mathbf{V}$, input data $\mathbf{x} \in \mathbf{V}$, and a set of classes $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K\}$.

- Discriminant Function $f(\mathbf{x})$

$$f : \mathbf{V} \to \mathcal{C}$$

- Discriminant Model $p(\mathcal{C}_k \mid \mathbf{x})$, then use decision theory
- Generative Model $p(\mathbf{x}, \mathcal{C}_k)$, then use decision theory

Probability/Density Estimation

- Maximum Likelihood (ML)

$$\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D} \,|\, \boldsymbol{\theta})$$

- Maximum a Posteriori (MAP)

$$\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \,|\, \mathcal{D}) \propto p(\mathcal{D} \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})$$

- Bayesian

$$p(\boldsymbol{\theta} \,|\, \mathcal{D}^{(n)}) \propto p(\mathbf{x}^{(n)} \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \,|\, \mathcal{D}^{(n-1)})$$

Being Bayesian

- Bayesian = maintaining a distribution
  - for any quantity of interest (e.g. position)
  - Bayesian parameter estimation
- Key idea: robust to overfitting
  - maintains varying strength of belief in multiple hypothesis
- In the limit of infinite data: ML = Bayesian
  - Data ‘swamps’ prior
  - Can you explain why?

The Role of Training Data

- Parametric Methods : Learn the model parameter from the training data, then discard training data.
- Nonparametric methods: Use training data for prediction
  - Histogram method
  - k-nearest neighbours
  - Parzen probability density model: set of function centered on the data
- Kernel methods: Use linear combination of functions evaluated at the training data.

Linear Regression

- General Regression Setup
- Closed-form solution
- Maximum Likelihood and Least Squares
- Geometry of Least Squares
- Sequential Learning (on-line)
- Choice of basis function
- Regularisation
- Powerful with nonlinear feature mappings
- Bias-Variance Decomposition

(Chapter 3.1, 3.2, 3.3)

Bayesian Linear Regression

- Closed-form solution
- Predictive Distribution

$$p(t \mid x, \mathbf{x}, \mathbf{t}) = \int p(t, \mathbf{w} \mid x, \mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{w} = \int p(t \mid \mathbf{w}, x) \, p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}) \, \mathrm{d}\mathbf{w}$$

- Conjugate Prior
- Limitations of Linear Basis Function Models
- Curse of dimensionality

Linear Classification

- General Classification Setup
- Input space versus Feature space
- Binary and Multiclass Labels
- Fisher's Linear Discriminant
- Perceptron Algorithm (Discontinuous activation function)
- Maximum Likelihood solution

$$\boldsymbol{\theta}^{\star} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X}, \mathbf{t} \mid \boldsymbol{\theta})$$

- Naive Bayes : all features conditioned on the class are independent of each other

(Chapter 4)

Logistic Regression

- Smooth logistic sigmoid acting on a linear feature vector
- Compare to perceptron
- Error as negative log likelihood (cross-entropy error)
- Gradient of error is target deviation times basis function (linear)
- Laplace approximation

Flexible Basis Functions

- Neural Networks
- Multilayer Perceptron with differentiable activation function
- The basis functions can now adopt to the data.
- Weight space symmetries.
- Error Backpropagation.
- Regularisation in Neural Networks.

(Chapter 5)

Kernels

- Inner Product $\rightarrow$ Kernel
- Kernels are a kind of similarity measure
- Sparse Kernel Machines
- Support Vector Machines
- How do we get to the relevant data points?
- Overlapping class distribution
- Output are decisions, not posterior probabilities.
- Relevance Vector Machines: Bayesian Sparse Kernel technique for classification and regression.

(Chapter 6.1, 6.2 and 7)

Dimensionality Reduction - PCA

- Maximise Variance
- Find the eigenvectors of the covariance corresponding to the largest eigenvalues.
- PCA and Compression
- Data Standardisation
- Data Whitening

(Chapter 12.1)

Autoencoders

- For feedforward neural networks, multiple layers can be advantageous
- Multiple PCA layers is equivalent to one single PCA
- Want to minimise reconstruction error, add nonlinear hidden layer
- Undercomplete autoencoder - lossy compression
- Pre-training of supervised learning (with unlabelled data)
- Denoising autoencoder
- Overcomplete autoencoder - sparse representations

Sum-Product Algorithm

- Independence structure - Local computations
- Sum-Product Algorithm
- Message passing
- Distributive Law

(Chapter 8.4)

Mixture Models

- Joint Probability over observed variables does not longer factorise.
- Introduce discrete latent variables to model complex marginal distributions over the observed variables by simpler distributions over observed and latent variables.
- K-means clustering
- Data compression
- Mixture of Bernoulli
- Mixture of Gaussians

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})\, p(\mathbf{x} \mid \mathbf{z}) = \sum_{\mathbf{z}} \prod_{k=1}^{K} \pi_k^{z_k} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$
$$= \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

(Chapter 9.1, 9.2)

Expectation Maximisation (EM)

- Evaluate the responsibilities, then maximise the parameters.
- E step: Find $p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
- M step: Find $\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ where

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\theta})$$

- Kullback-Leibler Divergence

(Chapter 9.3, 9.4)

Sequential Data

- Stationary vs. Nonstationary Sequential Distributions
- Markov Model of order $M = 0, 1, \dots$
- State Space Model using latent variables
- Hidden Markov Model (HMM): Latent variables are discrete.
- Homogenuous HMM
- Left-to-right HMM
- Viterbi algorithm

(Chapter 13.1, 13.2)

Flavour of this course

- Formalise intuitions about problems
- Use language of mathematics to express models
- Geometry, vectors, linear algebra for reasoning
- Probabilistic models to capture uncertainty
- Calculus to identify good parameters
- Design and analysis of algorithms
- Numerical algorithms in python
- Understand the choices when designing machine learning methods

Data61 | CSIRO
The Australian National University

What we did not cover (in detail)

- Other learning paradigms
  - Neural Networks
  - Evolutionary Methods (e.g. Genetic Algorithms)
  - Frequent item mining
  - Expert Systems / Rule based learning
- Theory
  - Information Theory
  - Convex Optimisation
  - Generalised Linear Models
  - Dynamical Systems
  - Reinforcement Learning
  - Artificial Intelligence
- Applications
  - Natural Language Processing
  - Computer Vision
  - Computational Social Science
  - Robotics

Research questions are everywhere

$$f_\theta(x) : \mathcal{X} \to \mathcal{Y}$$

Ong & Walder
Data61 | CSIRO
The Australian National University