Contents

1 Introduction and Probability Theory

An important hypothesis of Machine Learning

Pre-existing data repositories contain a lot of potentially valuable knowledge

Definition: What is Learning

(Semi-) automatic extraction of **valid, novel, useful and comprehensible knowledge** - in the form of rules, regularities, patterns, constraints or models - from arbitrary sets of data.
The goal is to develop efficient and useful algorithm.

Contents this subject covers:

Foundations of statistical learning, linear models, non-linear bases, kernel approaches, neural networks, Bayesian learning, probabilistic graphical models (Bayes Nets, Markov Random Fields), cluster analysis, dimensionality reduction, regularization and model selection

1.1 Machine Learning Basics

1.1.1 Terminologies

- **Instance:** measurements about individual entities/objects: e.g. a loan application (data points)
- **Attribute (Feature, explanatory variable):** component of the instances: e.g. the applicant's salary, numerics, etc. (x_i)
- **Label (Reponse, dependent variable):** an outcome that is categorical, numbers, etc. (y)
- **Examples:** instance coupled with label, i.e. $< x_i, y >$
- **Models:** discovered relationship between attributes and/or label.

1.1.2 Supervised vs Unsupervised Learning

	Data	Model used for
Supervised Learning	Labelled	Predict labels on new instances
Unsupervised Learning	Unlabelled	Cluster related instances; Project to fewer dimensions; Understand attribute relationships

Evaluations: important for supervised learning

Evaluation principle: measure quality is problem-dependent

Step 1: pick an evaluation metric Accuracy, Contingency table, Precision-Recall, F1, ROC curves

step 2: procure an independent, labelled **test sets**

Step 3: Average the evaluation metric over the test sets.

Cross Validation especially when data is poor.

1.1.3 Probability Theory

Three types of models

$\hat{y} = f(x)$: regressions, best fitting parameters given a model, we model x.
$p(y|x)$: For a given x, we model y as a distribution (likelihood of y given x).
$p(x, y)$: we model (x, y) together, the probability of having (x, y).

Definition: Random Variable

A random variable X is a **numerical function** of outcome $X(\omega) \in R$

Discrete distribution

- Govern r.v. taking discrete values
- Described by **probability mass function** $p(x)$ which is $p(X = X)$
- $P(X \leq x) = \sum_{a=-\infty}^{x} p(a)$
- **Examples:** Bernoulli, Binomial, Multinomial, Poisson

Continuous distribution

- Govern real-valued r.v.
- Described by **probability density function** $p(x)$ which is $p(X = X)$
- $P(X \leq x) = \int_{\infty}^{x} p(a)da$ (Sum from the very left)
- **Examples:** Uniform, Normal, Laplace, Gamma, Beta, Dirichlet

Definition: Bayes' Theorem

In terms of events A, B:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \Leftrightarrow P(A|B)P(B) = P(B|A)P(A)$$

Bayesian statistical inference makes heavy use of:

- **Marginals:** probabilities of individual variables
- **Marginalisation:** summing away all but r.v.'s of interest (reduce the number of variables/distribution)

2 Statistical Schools of Thought

2.1 Frequentist statistics

Wherein unknown model parameters are treated as having fixed but unknown values.

The problem that we intend to solve

Given: $X_1, X_2, ..., X_n$ drawn i.i.d from some unknown distribution
Want to: identify unknown distribution
Approach: Parametric Approach

- **Some model:** parameterised by parameter sets θ
- **Point estimates:** point estimates $\hat{(\theta)}$ a function or statistics of data.

Definition: Bias-Variance Decomposition

- **Bias:** $B_\theta(\hat{\theta}) = E_\theta[(\hat{\theta}(X_1, ...X_n)] - \theta$
- **Variance:** $Var_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - E_\theta[\hat{\theta}])^2]$
- **Bias-variance decomposition of square loss:**

$$E_\theta[(\theta - \hat{\theta})^2] = [B(\theta)]^2 + Var_\theta(\hat{\theta})$$

What we really care is the squared loss, which contains both bias and variance. (Notice that empirically, we care about the expected loss, since we don't know the true distribution, we use the distribution of the sample to approximate).

Asymptotic properties

Consistency: $\hat{\theta}(x_1, ...X_N)$ converges to true θ as $n \to \infty$ (i.e., if we increase the sample size to infinity, does the estimated distribution converge to the true distribution?)
Efficiency: asymptotic variance is as small as possible.

The Approaches: Maximum Likelihood Estimation (MLE)

Algorithm 1 Maximum Likelihood Estimation (MLE)

1: We have a set of data $(X_1, ..., X_n)$

2: We propose a distribution, p_θ, which is assumed to have generated the data (i.e. we assume that the data is generated by using this distribution function)

3: Express the likelihood of the data:

$$\prod_{i=1}^{n} p_\theta(X_i)$$

4: Apply log trick

5: Optimise to find the best (most likely) parameters θ: two ways to do that

6: (1) F.O.C: take partial derivatives of log likelihood w.r.t $\hat{\theta}$

7: (2) iterative processes, Newton method etc.

2.2 Bayesian Statistics

Wherein unknown model parameters have associated distributions reflecting prior belief.

Key idea: **Probabilities** $\Leftrightarrow$ **beliefs**

Bayesian Machine Learning:

Step 1: Start with prior $P(\theta)$ and likelihood $P(X|\theta)$

step 2: Observe data $X = x$

Step 3: Update prior to posterior $P(\theta|X = x)$

Definition: Bayes' Theorem

In terms of events A, B:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \Leftrightarrow P(A|B)P(B) = P(B|A)P(A)$$

Bayesian statistical inference makes heavy use of:

- **Marginals:** probabilities of individual variables
- **Marginalisation:** summing away all but r.v.'s of interest (reduce the number of unwanted variables/distribution)

$$P(X = x) = \sum_t P(X = x, \theta = t)$$