EECS 445

Introduction to Machine Learning

Honglak Lee - Fall 2015

Contributors: Max Smith

Latest revision: May 1, 2015

Contents

Abstract

Theory and implementation of state-of-the-art machine learning algorithms for large-scale real-world applications. Topics include supervised learning (regression, classification, kernel methods, neural networks, and regularization) and unsupervised learning (clustering, density estimation, and dimensionality reduction).

1 Readings

1.1 Probability Distributions

Definition 1.1 (Binary Variable)**.** Single variable that can take on either 1, or 0. $x \in \{0, 1\}$

We denote μ $(0 \leq \mu \leq 1)$ to be the probability that the random binary variable $x = 1$

$$p(x = 1|\mu) = \mu$$

$$p(x = 0|\mu) = 1 - \mu$$

Definition 1.2 (Bernoulli Distribution)**.** Probability distribution of the binary variable x, where μ is the probability $x = 1$.

$$\text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}$$

The distribution has the following properties:

- $\text{E}(x) = \mu$
- $\text{Var}(x) = \mu(1-\mu)$
- $\mathcal{D} = \{x_1, \ldots, x_N\} \to p(\mathcal{D}|\mu) = \Pi_{n=1}^{N} p(x_n|\mu)$
- Maximum likelihood estimator: $\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{numOfOnes}{sampleSize}$ (aka. sample mean)

Definition 1.3 (Binomial Distribution)**.** Distribution of m observations of $x = 1$, given a sample size of N.

$$\text{Bin}(m|N, \mu = {}^{N}_{m}\mu^m(1-\mu)^{N-m}$$

- $\text{E}(m) = N\mu$
- $\text{Var}(m) = N\mu(1-\mu)$

The Beta Distribution

In order to develop a Bayesian treatment for fitting data sets, we will introduce a prior distribution $p(\mu)$.

Definition 1.4 (Conjugacy)**.** when the prior and posterior distributions belong to the same family.

Definition 1.5 (Beta Distribution)**.**

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

Where $\Gamma(x)$ is the gamma function. The distribution has the following properties:

- $\text{E}(\mu) = \frac{a}{a+b}$
- $\text{Var}(\mu) = \frac{ab}{(a+b)^2(a+b+1)}$

- conjugacy
- $a \to \infty || b \to \infty \to$ variance
  $to 0$

Conjugacy can be shown by the distribution by the likelihood function (binomial):

$$p(\mu|m, l, a, b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1}$$

Normalized to:

$$p(\mu|m, l, a, b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1}$$

Definition 1.6 (Hyperparameters)**.** parameters that control the distribution of the regular parameters.

Definition 1.7 (Sequential Approach)**.** method of learning where you make use of an observation one at a time, or in small batches, and then discard them before the next observatiosn are used. (Can be shown with a Beta, where observing $x = 1 \to a++, x = 0 \to b++$, then normalizing)

Remark 1.1. For a finite data set, the posterior mean for μ always lies between the prior mean and the maximum likelihood estimate.

Remark 1.2. A general property of Bayesian learning is when we observe more and more data the uncertainty of the posterior distribution will steadily decrease.

More information and examples of probability distributions can be found in Appendix B of Bishop's 'Pattern Recognition and Machine Learning.'

1.2 Linear Models for Regression