Indian Institute of Technology, Kanpur
Probabilistic Machine Learning and Inference, 18–19

Instructors: Piyush Rai
Authors: Gurpreet Singh
Date: Decemenber 1, 2017

SCRIBE

1

Introduction to Probabilistic Machine Learning

Generally, the data we observe has some noise. Sometimes, we need to model this error or uncertainty explicitly. Often, we also need probabilistic predictions (fraudulent transaction). For this we need probabilistic modelling of our problem.

There could also be uncertainties in the estimated model parameters, or the predictions itself can be uncertain. To model these uncertainties, we require a probabilitic approach.

1. Probabilistic Modelling of Data

We generally assume the data points to be independent of each other, and dependent on only unknown paramters (say $\boldsymbol{\theta}$) *i.e.* we ssume data $\mathbf{X} = \{\mathbf{x}^1, \mathbf{x}^2 \dots \mathbf{x}^N\}$ to be generated from a probabilistic model with an unknown parameters $\boldsymbol{\theta}$

$$\mathbf{x}^1, \mathbf{x}^2 \dots \mathbf{x}^N \quad \sim \quad \mathbb{P}\left[\mathbf{X} \mid \boldsymbol{\theta}\right]$$

Since the points are assumed to be independent, we can aslo say

$$\mathbb{P}\left[\mathbf{X} \mid \boldsymbol{\theta}\right] \quad = \quad \prod_{n=1}^{N} \mathbb{P}\left[\mathbf{x}^n \mid \boldsymbol{\theta}\right]$$

1.1 Representing Probabilitic Models

We generally use a simplistic "plate-notation" graphical model to represent probabilistic models.

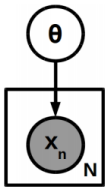

Figure 1: Plate-Notation Graphical Model

In the plate-notation, the shaded nodes represent known or observed values, whereas unshaded nodes represent unknown or learned variables, and the arrows show dependency.

In the example in figure 1, $\mathbf{X} = \{\mathbf{x}^1, \mathbf{x}^2 \dots \mathbf{x}^N\}$ is known, and $\boldsymbol{\theta}$ is learned. Also, $\mathbf{X}$ is dependent on $\boldsymbol{\theta}$.

Note. Partially shaded nodes represent partially observed or missing values. For example, in semi-supervised learning

Any node that we are uncertain about (can also be the observed nodes) is modelled using a probability distribution and these nodes become the random variables of the model.

The full model is specified via a joint probability distribution over all random variables.

1.2 Parameter Estimation

Specification of probabilistic models requires two key ingredients – Likelihood and Prior

Definition 1.1 (Likelihood Function)**.** This is the observation model that specifies how data is generated. This measures the data fit w.r.t. to the given parameters $\boldsymbol{\theta}$ and is represented by the conditional property $\mathbb{P}\left[\,\mathbf{X} \mid \boldsymbol{\theta}\,\right]$

Definition 1.2. Prior Distribution specifies how likely different parameter values are a priori. This also acts as a regularizer to the model and is represented by the prior probability $\mathbb{P}\left[\,\boldsymbol{\theta}\,\right]$

For parameter estimation, we have two methods, point estimate or Bayesian Inference.

A single point estimate is represented as the simple maximization problem

$$\widehat{\boldsymbol{\theta}}_{MLE} \quad = \quad \arg\max_{\boldsymbol{\theta}} \; \mathbb{P}\left[\,\mathbf{X} \mid \boldsymbol{\theta}\,\right]$$

This is known as the Maximum Likelihood Estimate (MLE). We can also have a regularized version, by maximising the posterior of the parameter, instead of the Likelihood Function. This is known as the Maximum-a-Posteriori (MAP) Estimate.

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{MLE} \quad &= \quad \arg\max_{\boldsymbol{\theta}} \; \mathbb{P}\left[\,\boldsymbol{\theta} \mid \mathbf{X}\,\right] \\ &= \quad \arg\max_{\boldsymbol{\theta}} \; \mathbb{P}\left[\,\mathbf{X} \mid \boldsymbol{\theta}\,\right] \mathbb{P}\left[\,\boldsymbol{\theta}\,\right] \end{aligned}$$

However, we cannot measure the uncertainty in the estimated parameter value. Hence, we use Bayesian Inference. We can estimate the full posterior distribution over $\boldsymbol{\theta}$ to get the uncertainty.

$$\mathbb{P}\left[\,\boldsymbol{\theta} \mid \mathbf{X}\,\right] \quad = \quad \frac{\mathbb{P}\left[\,\mathbf{X} \mid \boldsymbol{\theta}\,\right] \mathbb{P}\left[\,\boldsymbol{\theta}\,\right]}{\mathbb{P}\left[\,\mathbf{X}\,\right]}$$

We can also use Bayesian Inference in an online setting, where the old posterior can act as the new prior. Hence, our belief about $\boldsymbol{\theta}$ keeps getting updated as wee see more and more data.

2. Generative Modelling

Generative models are typically probabilistic, specifying a joint probability distribution over observation and target (label) values. A conditional distribution can be formed from a generative model through Bayes' rule.

Each data point $\mathbf{x}^n$ is associated with a latent variable $\mathbf{z}^n$, which is essentially a compact representation or encoding of the data point. For example, in case of Gaussian Mixture Models, the latent variable is the cluster assignment. The latent variables can also be used to infer missing data or *relevance* of a data point (hence generative).

Generative models are used in many problems (mostly unsupervised) such as GMM, Probabilistic PCA, Deep Generative Models, etc.