Pattern Recognition and Machine Learning Chapter 2 Summary

CSCI-GA 3033 Bayesian Machine Learning
New York University
Yilun Kuang

Sep 19, 2022

2.0 Introduction

We're interested in the problem of density estimation, i.e. modeling $p(\mathbf{x})$ given finite $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ observations. One example is *Language Modeling*, where $X = (x_1, ..., x_T)$ is a sequence of discrete tokens and we would like to assign a probability measure $p(X)$ for all possible sequences $X \in \mathcal{X}$. This is a fundamentally ill-posed problem as the possible set of data generating probability distributions is infinitely large.

This chapter presents examples of simple density estimation using parametric distributions and non-parametric methods. For parametric distributions, frequentist estimates the parameters by optimizing the likelihood function, and the Bayesian approach chooses conjugate priors and uses Bayes Theorem to induce a posterior measure over the parameters.

2.1 - 2.2 Binary and Multinomial Variables

Frequentist's MLE estimation of parameters of the Binomial distribution can easily overfit. From a Bayesian perspective, the product of Beta prior and Binomial likelihood is a Beta posterior, which incorporates prior information and data and is less likely to overfit. To make prediction, we can invoke the predictive distribution. In the asymptotic infinite data regime, MLE estimation is equivalent to Bayesian predictive solution and the posterior variance shrink. This is consistent with the Bayesian Inference approach in perceptual psychologies and neuroscience. Similarly, a Dirichlet prior multiplying a Multinomial likelihood leads to a Dirichlet posterior.

2.3 Gaussian Variables

2.3.0 Limitation of Gaussians

Gaussian distribution as a density model is computationally expensive unless we choose diagonal or even isotropic covariances over general covariance matrices, but then we're limited by the approximation power. Gaussian is also unimodal, unless we invoke mixtures of Gaussians or other techniques like Probabilistic Graphical Models.

2.3.1 - 2.3.3 Marginal, Conditional Gaussian, and Bayes Theorem for Gaussian

If $p(\mathbf{x}_a, \mathbf{x}_b)$ is Gaussian, then $p(\mathbf{x}_a|\mathbf{x}_b)$ and $p(\mathbf{x}_a)$ are also Gaussian. Some of the techniques here include *completing the squares* and *Schur complement*. On Page 93 there is a summary of useful Gaussian results under the Bayes Theorem.

Marginal and Conditional Gaussians

Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \tag{2.113}$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \tag{2.114}$$

the marginal distribution of $\mathbf{y}$ and the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \tag{2.115}$$
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \tag{2.116}$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}. \tag{2.117}$$

Figure 1: Bayes Theorem for Gaussian

2.3.4 - 2.3.6 Density Estimation of Gaussian

Frequentist MLE estimation of Gaussian parameters can be extended to sequential settings with *Robbins-Monro* algorithm. Finding the MLE solutions corresponds to finding the root of a regression function using the *Robbins-Monro* algorithm.

From a Bayesian perspective, for a univariate / multivariate Gaussian, we have the following settings:

- If variance / covariance is known and the goal is to infer the mean
  - Conjugate Prior (Gaussian) * Likelihood (Gaussian) = Posterior (Gaussian)
  - In the infinite asymptotic data regime, posterior mean & variance / covariance is equivalent to MLE estimate of mean & variance / covariance in the frequentist approach
- If the mean is known and the goal is to infer precision (inverse of variance / covariance)
  - Conjugate Prior (Gamma / Wishart) * Likelihood (Gaussian) = Posterior (Gamma)
- If the mean is known and the goal is to infer variance / covariance
  - Conjugate Prior (Inverse Gamma / Inverse Wishart) * Likelihood (Gaussian)
- If both the mean and the precision is unknown
  - Conjugate Prior (Gaussian-gamma / Gaussian-wishart) * Likelihood (Gaussian)

Sequentially, the learnt posterior can be used as the new prior for online learning.

2.3.7 - 2.3.9 Student's t-Distribution, *von Mises* Distribution, Mixture of Gaussians

If we multiply Gamma prior and the univariate Gaussian likelihood and marginalize out the precision, we get Student's t-distribution (for $\nu \to 1$, Cauchy; $\nu \to \infty$, Gaussian, i.e. equivalent to infinite mixture of Gaussians). T-Distribution is robust as it has heavy tails.

The *von Mises* / circular normal distribution is a periodic generalization of Gaussian. Mixture of *von Mises* distributions can be used to model periodic variables with multimodality.

The mixture of Gaussians is given by $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \boldsymbol{\Sigma}_k)$, where the posterior $p(k|\mathbf{x})$ is called "responsibilities". MLE for Mixture of Gaussians does not have a closed-form solution. We can solve the

MLE by Expectation-Maximization (EM) algorithm.

2.4 - 2.5 The Exponential Family and Nonparametric Method

The exponential family of distributions over $\mathbf{x}$ with parameters $\boldsymbol{\eta}$ is defined as

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x})\}$$

Bernoulli (with sigmoid jumping out), Multinomial (with softmax coming out), and Gaussian are all parts of the exponential family. For MLE estimation, to obtain $\boldsymbol{\eta}_{\text{ML}}$, it suffices to look at $\sum_n \mathbf{u}(\mathbf{x}_n)$, which is called the sufficient statistic of the distribution.

For now we only concerns frequentist nonparametric methods. The histogram approach causes discontinuities due to bin edges and suffers from the curse of dimensionality. For $p(\mathbf{x}) = \frac{K}{NV}$ as the density estimate, fixing K and determining V leads to the K-Nearest-Neighbor and fixing V and determining K leads to the Kernel Density Estimation. Both Kernel Density Estimation and K-Nearest-Neighbor are computationally expensive.

Miscellaneous

Gaussian

- For a single variable, the distribution that maximizes entropy is Gaussian.
- Central Limit Theorem
  - Binomial tends to Gaussian as $N \to \infty$, where N is the number of observations of the random binary variable x.
- We can change the basis of Gaussian covariance matrix to the eigenvector basis so that the multivariate normal factors into a product of independent univariate distribution.
- Second order moments for Gaussian
  - $\mathbb{E}[x^2]$ for univariate.
  - multivariate
    - $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mu\mu^\top + \Sigma$.

Exponential Family

This summary omits the 2.4.2-2.4.3 General Conjugate Priors and Non-informative Priors.