Density Estimation

Seungjin Choi

Department of Computer Science and Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 37673, Korea
seungjin@postech.ac.kr

Supervised vs Unsupervised Learning

Main goal of learning is to train probabilistic models from observed data which are given as $\mathcal{D} = \{(\boldsymbol{x}_t, \boldsymbol{y}_t)\}$ for supervised learning and $\mathcal{D} = \{\boldsymbol{x}_t\}$ for unsupervised learning.

- Supervised learning
  - Use input and target samples and learn a mapping from input to output under a probabilistic model
  - $\mathcal{D} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$
  - Assume a parameterized model $p(\boldsymbol{y}|\boldsymbol{x}, \theta) = \int p(\boldsymbol{y}, \boldsymbol{s}|\boldsymbol{x}, \theta) d\boldsymbol{s}$.
- Unsupervised learning
  - Take a set of input samples and fit a probabilistic model
  - $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$
  - Assume a parameterized model $p(\boldsymbol{x}|\theta) = \int p(\boldsymbol{x}, \boldsymbol{s}|\theta) d\boldsymbol{s}$.

Density Estimation

- The density estimation is the problem of modeling a probability density function $p(\boldsymbol{x})$, given a finite number of data points, $\{\boldsymbol{x}_t\}_{t=1}^{N}$ drawn from that density function.
- Approaches to density estimation
  - Parametric estimation
    - Assumes a specific functional form for density model
    - A number of parameters are optimized by fitting the model to the data set
  - Nonparametric estimation
    - No specific functional form is assumed
    - Allows the form of the density to be determined entirely by the data
  - Semi-parametric estimation
    - Tries to achieve the best of both worlds
    - Flexible models (feedforward neural nets and mixture models)

Parametric Methods

- Represent $p(\boldsymbol{x})$ in terms of a specific functional form which has a number of adjustable parameters.
- For example, multivariate Gaussian distribution that is of the form

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{1}{D}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right\}.$$

- Methods for parameter estimation
  - Maximum likelihood estimation
  - MAP estimation
  - Bayesian inference

Why Gaussian Distribution?

- Simple analytical properties.
- Completely characterized by mean and covariance.
- Central limit theorem.
- The distribution is again normal after a nonsingular linear transform.
- Marginal density is also normal and conditional density is also normal.
- There exists a linear transform which diagonalizes the covariance matrix (whitening, data sphering).
- Has maximum entropy, given values of the mean and the covariance matrix.

Maximum Likelihood Estimation (MLE)

- The likelihood function is nothing but a parameterized density $p(\boldsymbol{x}|\theta)$ that is used to model a set of data $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ which are assumed to be drawn independently from $p(\boldsymbol{x}|\theta)$:

$$p(\mathcal{X}|\theta) = \prod_{t=1}^{N} p(\boldsymbol{x}_t|\theta).$$

- Maximum likelihood seeks to find the optimum values for the parameters by maximizing a likelihood function form the training data.
- The log-likelihood is given by

$$\mathcal{L}(\theta) = \sum_{t=1}^{N} \log p(\boldsymbol{x}_t|\theta).$$

- ML finds $\widehat{\theta}_{ML}$:

$$\widehat{\theta}_{ML} = \arg\max_{\theta} \mathcal{L}(\theta).$$

Density Estimation: MLE vs Bayesian

Given a set of data, $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$, density estimation involves estimating a distribution from which observed data points $\boldsymbol{x}_t$ are drawn.

Maximum Likelihood

- Model $\boldsymbol{x}_t|\theta \sim p(\cdot|\theta)$.
- Find MLE:

$$\begin{aligned}\theta_{ML} &= \arg\max_\theta \log p(\boldsymbol{X}|\theta) \\ &= \arg\max_\theta \sum_{t=1}^{N} \log p(\boldsymbol{x}_t|\theta).\end{aligned}$$

- Prediction is done by

$$p(\boldsymbol{x}_*|\theta_{ML}).$$

Bayesian

- Model $\boldsymbol{x}_t|\theta \sim p(\cdot|\theta)$.
- Prior over parameters: $p(\theta)$.
- Posterior over parameters

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})}.$$

- Prediction is done by

$$p(\boldsymbol{x}_*|\boldsymbol{X}) = \int p(\boldsymbol{x}_*|\theta)p(\theta|\boldsymbol{X})d\theta.$$

MLE: Kullback Matching Perspective

Suppose that we are given a set of data, $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$ drawn from an underlying distribution $p(\boldsymbol{x})$.

- Empirical distribution: $\widetilde{p}(\boldsymbol{x}) = \frac{1}{N}\sum_{t=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_t)$.
- Model: $p(\boldsymbol{x}|\theta)$.

Fit the model $p(\boldsymbol{x}|\theta)$ to data $\boldsymbol{X}$:

$$\begin{aligned}
\arg\min_{\theta} KL[\widetilde{p}(\boldsymbol{x})||p(\boldsymbol{x}|\theta)] &= \arg\min_{\theta} \int \widetilde{p}(\boldsymbol{x}) \log \frac{\widetilde{p}(\boldsymbol{x})}{p(\boldsymbol{x}|\theta)} d\boldsymbol{x} \\
&= \arg\min_{\theta} \Big[-H(\widetilde{p}) - \int \widetilde{p}(\boldsymbol{x}) \log p(\boldsymbol{x}|\theta) d\boldsymbol{x} \Big],
\end{aligned}$$

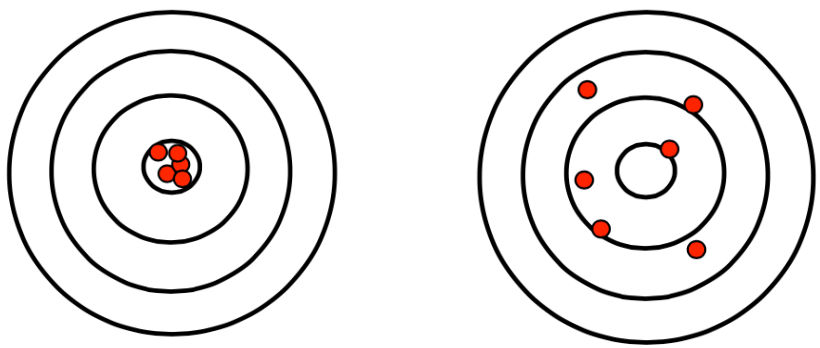
leading to

$$\begin{aligned}
\arg\max_{\theta} \mathbb{E}_{\widetilde{p}} \log p(\boldsymbol{x}|\boldsymbol{\theta}) &= \arg\max_{\theta} \frac{1}{N} \int \sum_{t=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_t) \log p(\boldsymbol{x}|\theta) d\boldsymbol{x} \\
&= \arg\max_{\theta} \frac{1}{N} \sum_{t=1}^{N} \log p(\boldsymbol{x}_t|\boldsymbol{\theta}).
\end{aligned}$$

Estimation

- Estimator: Statistic whose calculated value is used to estimate model parameter θ
- Estimate: A particular realization of an estimator, $\widehat{\theta}$.

- Good estimators are:
  - Consistent: $\lim_{n\to\infty} P(|\widehat{\theta} - \theta_*| > \epsilon) = 0$.
  - Unbiased: $\mathbb{E}_{p(\boldsymbol{X}|\theta)}\left[\widehat{\theta}\right] = \theta_*$.



Parameter Estimation: An Example

Suppose that we wish to estimate θ from its noisy observations $x_t = \theta + \epsilon_t$ for $t = 1, \ldots, N$, where $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$.

- Estimator: Take the first sample only

$$\widehat{\theta} = x_1.$$

- Mean and variance:

$$\mathbb{E}\left[\widehat{\theta}\right] = \theta, \quad \text{var}(\widehat{\theta}) = \sigma^2.$$

- Estimator: Take averaging

$$\overline{\theta} = \frac{1}{N}\sum_{t=1}^{N} x_t.$$

- Mean and variance:

$$\mathbb{E}\left[\overline{\theta}\right] = \theta, \quad \text{var}(\overline{\theta}) = \frac{\sigma^2}{N}.$$

Both estimators are unbiased, but $\text{var}(\overline{\theta}) \leq \text{var}(\widehat{\theta})$.

It turns out that $\overline{\theta} = \theta_{ML}$.

Maximum Likelihood: Presence of Latent Variables

In the presence of latent variable $\boldsymbol{s}$, a parameterized model is given by

$$p(\boldsymbol{x}|\theta) = \int p(\boldsymbol{x}, \boldsymbol{s}|\theta) d\boldsymbol{s}.$$

MLE is determined by

$$\arg\max_{\theta} \log \int p(\boldsymbol{x}, \boldsymbol{s}|\theta) d\boldsymbol{s}.$$

Not easy to find MLE $\Rightarrow$ EM.

An Example of ML: Univariate Normal

Suppose that our parameterized density $p(x|\theta)$ is given by

$$p(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}.$$

Let $\theta_1 = \mu$ and $\theta_2 = \sigma^2$ are unknown parameters.
We consider the log-likelihood $\mathcal{L}(\theta)$ given by

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_{t=1}^{N} \log p(x_t|\theta) \\ &= \sum_{t=1}^{N} \left[-\frac{1}{2}\log 2\pi\theta_2 - \frac{1}{2\theta_2}(x_t-\theta_1)^2\right].\end{aligned}$$

An Example of ML: Univariate Normal (Cont'd)

We find stationary points, solving $\nabla_\theta \log p(\mathcal{X}|\theta) = 0$, given by

$$\begin{aligned} \sum_{t=1}^{N} \frac{1}{\theta_2}(x_t - \theta_1) &= 0, \\ -\sum_{t=1}^{N} \frac{1}{2\theta_2} + \sum_{t=1}^{N} \frac{(x_t - \theta_1)^2}{2\theta_2^2} &= 0. \end{aligned}$$

These lead to maximum likelihood estimates:

$$\begin{aligned} \widehat{\theta}_1 &= \frac{1}{N}\sum_{t=1}^{N} x_t, \quad \text{(sample mean)} \\ \widehat{\theta}_2 &= \frac{1}{N}\sum_{t=1}^{N} \left(x_t - \widehat{\theta}_1\right)^2. \quad \text{(sample variance)} \end{aligned}$$

MAP Estimation

$$
\begin{aligned}
\widehat{\theta}_{MAP} &= \arg\max_{\theta} p(\theta|\boldsymbol{x}) \\
&= \arg\max_{\theta} \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})} \\
&= \arg\max_{\theta} p(\boldsymbol{x}|\theta)p(\theta) \\
&= \arg\max_{\theta} \left[\log p(\boldsymbol{x}|\theta) + \log p(\theta)\right].
\end{aligned}
$$

- The prior $\log p(\theta)$ plays a critical role in protecting against overfitting.
- If our belief says the function should be smooth, then the prior plays like an regularizer (which penalizes too complex models).

An Example of MAP Estimation: Univariate Normal

Assume $x \sim \mathcal{N}(\mu, 1)$. Use a prior $p(\mu) \sim \mathcal{N}(0, \alpha^2)$.
Then we have

$$\begin{aligned} \mathcal{L} &= \log p(\mathcal{X}|\theta) + \log p(\theta) \\ &\propto -\frac{1}{2}\sum_{t=1}^{N}(x_t - \mu)^2 - \frac{1}{2\alpha^2}\mu^2. \end{aligned}$$

It follows from $\frac{\partial \mathcal{L}}{\partial \mu} = 0$ that

$$\widehat{\mu}_{MAP} = \frac{1}{\left(N + \frac{1}{\alpha^2}\right)}\sum_{t=1}^{N} x_t.$$

Remarks from this Example

- For $N \gg \frac{1}{\alpha^2}$ (the influence of the prior is negligible), we have

$$\widehat{\mu}_{MAP} \longrightarrow \widehat{\mu}_{ML} = \frac{1}{N} \sum_{t=1}^{N} x_t$$

- For very strong belief in the prior, i.e., $\frac{1}{\alpha^2} \gg N$, we have

$$\widehat{\mu}_{MAP} \longrightarrow 0.$$

  If few data points are available, the prior will bias the estimate towards the priori expected value.

Bayesian Inference

- A Bayesian considers θ as a random variable.
- A Bayesian wants to know how his prior knowledge of the random variable θ changes in the light of the new observations $\boldsymbol{d}$, where $\boldsymbol{d} = (\boldsymbol{x}, \boldsymbol{y})$ in the case of supervised learning and $\boldsymbol{d} = \boldsymbol{x}$ in the case of unsupervised learning.
- Need to calculate the posterior distribution

$$p(\theta|\boldsymbol{d}) = \frac{\overbrace{p(\boldsymbol{d}|\theta)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{\int p(\boldsymbol{d}|\theta)p(\theta)d\theta}_{\text{marginal likelihood}}}.$$

- In general, the marginal likelihood (or evidence) is hard to compute.

Bayesian Inference: Predictive Distribution

- The unsupervised Bayesian would want to calculate the probability of a new data point $\boldsymbol{x}$, given the data $\mathcal{D}$,

$$\begin{aligned} p(\boldsymbol{x}|\mathcal{D}) &= \int p(\boldsymbol{x}, \theta|\mathcal{D}) d\theta \\ &= \int p(\boldsymbol{x}|\theta, \mathcal{D}) p(\theta|\mathcal{D}) d\theta \\ &= \int p(\boldsymbol{x}|\theta) p(\theta|\mathcal{D}) d\theta. \end{aligned}$$

- The supervised Bayesian would want to calculate the probability over target values, given an input data point and the previous data points,

$$p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{D}) = \int p(\boldsymbol{y}|\boldsymbol{x}, \theta) p(\theta|\mathcal{D}) d\theta.$$

- Bayesian approach performs a weighted average over all values of θ, instead of choosing a specific value for θ.

Bayesian Inference: Posterior Calculation

The posterior distribution of θ is updated using Bayes rule, where the likelihood is given by $p(\mathcal{D}|\theta) = \prod_{t=1}^{N} p(\boldsymbol{x}_t|\theta)$:

$$\begin{aligned} p(\theta|\mathcal{D}) &= \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \\ &= \frac{p(\theta)p(\mathcal{D}|\theta)}{\int p(\mathcal{D}|\theta')p(\theta')d\theta'} \\ &= \frac{p(\theta)\prod_{t=1}^{N} p(\boldsymbol{x}_t|\theta)}{\int p(\theta')\prod_{t=1}^{N} p(\boldsymbol{x}_t|\theta')d\theta'}. \end{aligned}$$

Conjugate prior: A prior $p(\theta)$ which gives rise to a posterior $p(\theta|\mathcal{D})$ having the same function form, given $p(\mathcal{D}|\theta)$.

Bayesian Inference: A Few Remarks

- Never actually estimate a value of θ.
- Instead, determine the posterior density over all values for θ and use it to integrate over all possible values of θ.
- Approximation inference
  - Laplace approximation
  - Variational Bayes
  - Markov chain Monte Carlo (MCMC)

Bayesian Inference: An Example

Suppose $x \sim \mathcal{N}(\mu, \sigma^2)$ where σ^2 is assumed to be known.
Find the mean μ, given a set of data points $\{x_t\}$.
Assume that the prior for μ to be Gaussian,

$$p_0(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right\}.$$

Observing a set of N data points, we calculate the posterior

$$p(\mu|\mathcal{D}) = \frac{p_0(\mu)}{p(\mathcal{D})} \prod_{t=1}^{N} p(x_t|\mu),$$

where

$$p(x_t|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_t - \mu)^2\right\}.$$

After tedious calculations, we have

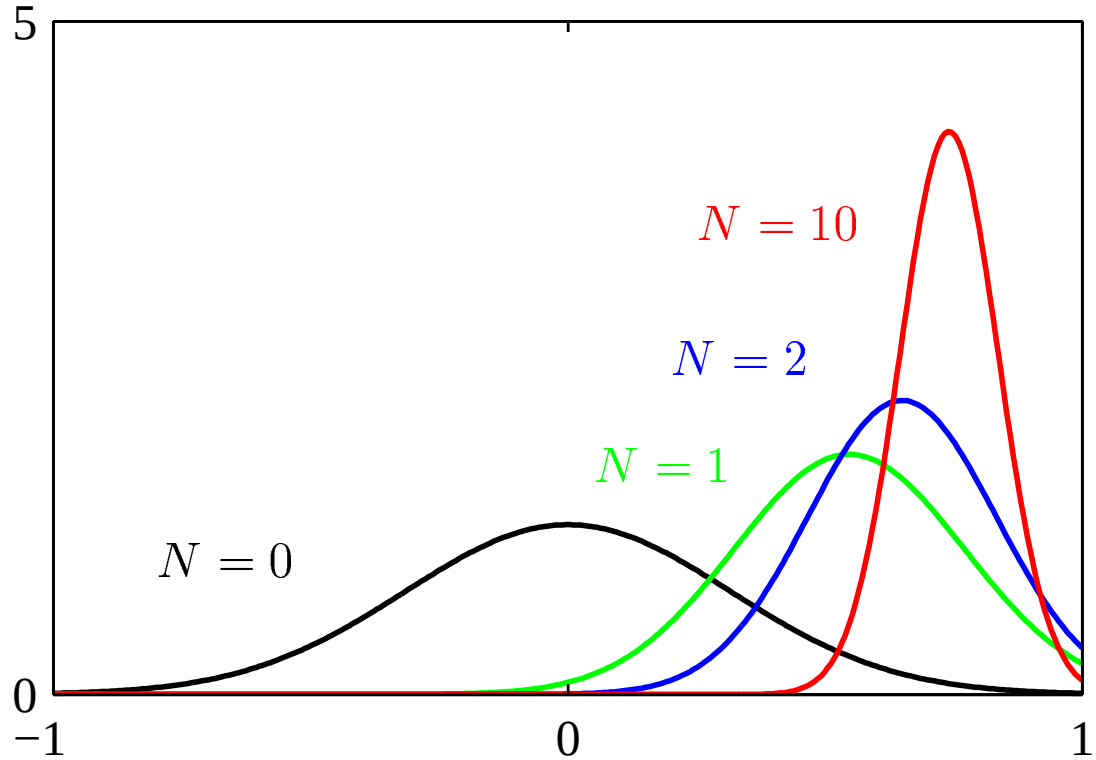
$$p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi\widetilde{\sigma}^2}} \exp\left\{-\frac{1}{2\widetilde{\sigma}^2}(\mu - \widetilde{\mu})^2\right\},$$

where

$$\begin{aligned}
\widetilde{\mu} &= \left(\frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\right)\mu_{ML} + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0, \\
\widetilde{\sigma}^2 &= \frac{\sigma_0^2\sigma^2}{N\sigma_0^2 + \sigma^2}, \qquad \frac{1}{\widetilde{\sigma}^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \text{ (precision)}.
\end{aligned}$$

- When $N = 0$, $\widetilde{\mu}$ reduces to the prior mean and $\widetilde{\sigma}^2$ does to the prior variance, as expected.
- As $N \to \infty$, the posterior mean is given by the ML solution and the posterior variance goes to 0, leading that the posterior distribution becomes infinitely peaked around the ML solution.

The data points are generated from a Gaussian of mean 0.8 and variance 0.1 and the prior is chosen to have mean 0. The posterior distribution is shown for increasing numbers N of data points.



Kernel Density Estimation: Nonparametric Approach

Place a kernel on each data point and compute an average to estimate the probability distribution of $\boldsymbol{x}$, given a set of data points $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$:

$$\begin{aligned} \widehat{p}(\boldsymbol{x}) &= \frac{1}{N} \sum_{t=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_t, \lambda_x) \\ &= \frac{1}{N} \sum_{t=1}^{N} \frac{1}{Z_x} \exp\left\{-\lambda_x \|\boldsymbol{x} - \boldsymbol{x}_t\|^2\right\}. \end{aligned}$$