Density Estimation (Lecture 2.2)

Alexander J. Smola
Alex.Smola@anu.edu.au

Machine Learning Program
National ICT Australia
RSISE, The Australian National University

This Week's Topics

- Probability Theory Basics
- Maximum Likelihood
- Estimators and Efficiency
- Priors
- Exponential Family

NATIONAL ICT AUSTRALIA

Estimators

Formalizing the Inference Process

Given some data X drawn from a distribution $p(x;\theta)$, estimate the parameter θ via the mapping $\hat{\theta}(X)$.

Unbiased Estimator

An unbiased estimator $\hat{\theta}(X)$ of the parameters θ of the distribution $p(x;\theta)$ satisfies

$$\mathbf{E}_{X\sim p(X;\theta)}[\hat{\theta}(X)] = \theta$$

Theorem

The Maximum-Likelihood Estimator

$$\hat{\theta}(X) := \operatorname*{argmax}_{\theta} p(X;\theta)$$

is asymptotically unbiased.

Warning: MLE need not give good results necessarily. Recall estimating probabilities for a dice.

NATIONAL ICT AUSTRALIA

Example: Normal Distribution

Density Model

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

Log-Likelihood

$$-\log p(X;\mu,\Sigma) = \sum_{i=1}^{m} \frac{n}{2}\log 2\pi + \frac{1}{2}\log|\Sigma| + \frac{1}{2}(x_i-\mu)^\top\Sigma^{-1}(x$$

$$\partial_\mu -\log p(X;\mu,\Sigma) = \sum_{i=1}^{m} \Sigma^{-1}(\mu - x_i)$$

$$\partial_\mu -\log p(X;\mu,\Sigma) = \sum_{i=1}^{m} \frac{1}{2}\Sigma^{-1} - \frac{1}{2}\Sigma^{-1}(x_i-\mu)(x_i-\mu)^\top\Sigma^{-1}$$

Hence $\mu = \frac{1}{m}\sum_{i=1}^m x_i$ and $\Sigma = \frac{1}{m}\sum_{i=1}^m (x_i-\mu)(x_i-\mu)^\top$.

NATIONAL ICT AUSTRALIA

Fisher Information and Efficiency

Fisher Score

$$V_\theta(x) := \partial_\theta \log p(x;\theta)$$

This tells us the influence of x on estimating θ. Its expected value vanishes, since

$$\begin{aligned}\mathbf{E}\left[\partial_\theta \log p(X;\theta)\right] &= \int p(X;\theta)\partial_\theta \log p(X;\theta) dX \\ &= \partial_\theta \int p(X;\theta) dX = 0.\end{aligned}$$

Fisher Information Matrix

It is the covariance matrix of the Fisher scores, that is

$$I := \mathrm{Cov}[V_\theta(x)]$$

NATIONAL ICT AUSTRALIA

Cramer Rao Theorem

Efficiency

Covariance of estimator $\hat{\theta}(X)$ rescaled by I:

$$e := \det \mathrm{Cov}[\hat{\theta}(X)]\mathrm{Cov}[\partial_\theta \log p(X;\theta)]$$

Theorem

The efficiency for unbiased estimators is never better (i.e. smaller) than 1. Equality is achieved for MLEs.

Proof (scalar case only)

By Cauchy-Schwartz we have

$$\left(\mathbf{E}_\theta\left[\left(V_\theta(X) - \mathbf{E}_\theta\left[V_\theta(X)\right]\right)\left(\hat{\theta}(X) - \mathbf{E}_\theta\left[\hat{\theta}(X)\right]\right)\right]\right)^2$$

$$\leq \mathbf{E}_\theta\left[\left(V_\theta(X) - \mathbf{E}_\theta\left[V_\theta(X)\right]\right)^2\right]\mathbf{E}_\theta\left[\left(\hat{\theta}(X) - \mathbf{E}_\theta\left[\hat{\theta}(X)\right]\right)^2\right] = IB$$

NATIONAL ICT AUSTRALIA

Cramer Rao Theorem

Proof

At the same time, $\mathbf{E}_\theta [V_\theta(X)] = 0$ implies that

$$\begin{aligned}
&\mathbf{E}_\theta \left[(V_\theta(X) - \mathbf{E}_\theta [V_\theta(X)]) \left(\hat{\theta}(X) - \mathbf{E}_\theta \left[\hat{\theta}(X) \right] \right) \right] \\
=&\mathbf{E}_\theta \left[V_\theta(X) \hat{\theta}(X) \right]^2 \\
=&\left(\int p(X|\theta) \partial_\theta p(X|\theta) \hat{\theta}(X) dX \right) \\
=&\partial_\theta \int p(X|\theta) \hat{\theta}(X) dX = \partial_\theta \theta = 1.
\end{aligned}$$

Cautionary Note

This does not imply that not a biased estimator might have lower variance.

NATIONAL ICT AUSTRALIA

The Exponential Family

Definition

A family of probability distributions which satisfy

$$p(x;\theta) = \exp(\langle \phi(x), \theta \rangle - g(\theta))$$

Details

- $\phi(x)$ is called the sufficient statistics of x.
- $\mathcal{X}$ is the domain out of which x is drawn ($x \in \mathcal{X}$).
- $g(\theta)$ is the log-partition function and it ensures that the distribution integrates out to 1.

$$g(\theta) = \log \int_{\mathcal{X}} \exp(\langle \phi(x), \theta \rangle) dx$$

NATIONAL ICT AUSTRALIA

Example: Binomial Distribution

Tossing coins

With probability p we have heads and with probability $1-p$ we see tails. So we have

$$p(x) = p^x(1-p)^{1-x} \text{ where } x \in \{0,1\} =: \mathcal{X}$$

Massaging the math

$$\begin{aligned} p(x) &= \exp \log p(x) \\ &= \exp\left(x \log p + (1-x)\log(1-p)\right) \\ &= \exp\Big(\big\langle \underbrace{(x, 1-x)}_{\phi(x)}, \underbrace{(\log p, \log(1-p))}_{\theta} \big\rangle\Big) \end{aligned}$$

The Normalization Once we relax the restriction on $\theta \in \mathbb{R}^2$ we need $g(\theta)$ which yields

$$g(\theta) = \log\left(e^{\theta_1} + e^{\theta_2}\right)$$

NATIONAL ICT AUSTRALIA

Example: Laplace Distribution

Atomic decay

At any time, with probability θdx an atom will decay in the time interval $[x, x + dx]$ if it still exists. Consulting your physics book tells us that this gives us the density

$$p(x) = \theta \exp(\theta x) \text{ where } x \in [0, \infty) =: \mathcal{X}$$

Massaging the math

$$p(x) = \exp\Big(\langle \underbrace{-x}_{\phi(x)}, \theta \rangle - \underbrace{-\log \theta}_{g(\theta)}\Big)$$

NATIONAL ICT AUSTRALIA

Example: Normal Distribution

Engineer's favorite

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \text{ where } x \in \mathbb{R} =: \mathcal{X}$$

Massaging the math

$$\begin{aligned} p(x) &= \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right) \\ &= \exp\Big(\langle \underbrace{(x, x^2)}_{\phi(x)}, \theta\rangle - \underbrace{\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)}_{g(\theta)}\Big) \end{aligned}$$

Finally we need to solve (μ, σ^2) for θ. Tedious algebra yields $\theta_2 := -\frac{1}{2}\sigma^{-2}$ and $\theta_1 := \mu\sigma^{-2}$. We have

$$g(\theta) = -\frac{1}{4}\theta_1^2\theta_2^{-1} + \frac{1}{2}\log 2\pi - \frac{1}{2}\log -2\theta_2$$

NATIONAL ICT AUSTRALIA

Example: Multinomial Distribution

Many discrete events

Assume that we have disjoint events $[1..n] =: \mathcal{X}$ which all may occur with a certain probability p_x.

Guessing the answer

Use the map $\phi : x \to e_x$, that is, e_x is an element of the canonical basis $(0, \ldots, 0, 1, 0, \ldots)$. This gives

$$p(x) = \exp(\langle e_x, \theta \rangle - g(\theta))$$

where the normalization is

$$g(\theta) = \log \sum_{i=1}^{n} \exp(\theta_i)$$

NATIONAL ICT AUSTRALIA

Benefits: Simple Estimation

Likelihood of a set: Given $X := \{x_1, \ldots, x_m\}$ we get

$$p(X;\theta) = \prod_{i=1}^{m} p(x_i;\theta) = \exp\left(\sum_{i=1}^{m} \langle \phi(x_i), \theta \rangle - m g(\theta)\right)$$
$$= \exp\left(m(\langle \mu, \theta \rangle - g(\theta))\right)$$

Here we set μ to the mean of the sufficient statistics

$$\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} \phi(x_i)$$

Maximum Likelihood

Derivative of the log-likelihood needs to vanish. This yields

$$\hat{\mu} - \partial_\theta g(\theta) = 0$$

NATIONAL ICT AUSTRALIA

Benefits: Log-partition function is nice

$g(\theta)$ generates moments:

$$\text{Recall: } g(\theta) = \log \int \exp(\langle \phi(x), \theta \rangle)$$

Taking the derivative wrt. θ we can see that

$$\partial_\theta g(\theta) = \mathbf{E}_{x \sim p(x;\theta)} [\phi(x)]$$
$$\partial_\theta^2 g(\theta) = \mathbf{Cov}_{x \sim p(x;\theta)} [\phi(x)]$$

. . . and so on for higher order moments . . .

Corollary

$g(\theta)$ is convex

Practical Benefit

Solving the problem $\mu = \partial_\theta g(\theta)$ becomes **easy.**

NATIONAL ICT AUSTRALIA

Application: Laplace distribution

Estimate the decay constant of an atom:

We use exponential family notation where

$$p(x;\theta) = \exp(\langle(-x),\theta\rangle - (-\log\theta))$$

Computing μ

Since $\phi(x) = -x$ all we need to do is **average over all decay times** that we observe.

Solving for Maximum Likelihood

The condition $\mu = \partial_\theta g(\theta)$ equates to $\mu = -\frac{1}{\theta}$. Solving for θ yields

$$\theta = -\frac{1}{\mu}$$

NATIONAL ICT AUSTRALIA

Benefits: Maximum Entropy Estimate

Entropy

Basically it's the number of bits needed to encode a random variable. It is defined as

$$H(p) = \int \log p(x) p(x) dx \text{ where we set } 0 \log 0 := 0$$

Maximum Entropy Density

The density $p(x)$ satisfying $\mathbf{E}[\phi(x)] \geq \eta$ with maximum entropy is $\exp(\langle \phi(x), \theta \rangle - g(\theta))$.

Corollary

The most vague density with a given variance is the Gaussian distribution.

Corollary

The most vague density with a given mean is the Laplacian distribution.

NATIONAL ICT AUSTRALIA

Using it

Observe Data

$x_1, \ldots, x_m$ drawn from distribution $p(x|\theta)$

Compute Likelihood

$$p(X|\theta) = \prod_{i=1}^{m} \exp(\langle \phi(x_i), \theta \rangle - g(\theta))$$

Maximize it

Take the negative log and minimize, which leads to

$$\partial_\theta g(\theta) = \frac{1}{m} \sum_{i=1}^{m} \phi(x_i)$$

This can be solved analytically or (whenever this is impossible or we are lazy) by Newton's method.

NATIONAL ICT AUSTRALIA

Priors

Problems with Maximum Likelihood

With not enough data, parameter estimates will be bad.

Prior to the rescue

Often we know where the solution should be. So we encode the latter by means of a prior $p(\theta)$.

Normal Prior

Simply set $p(\theta) \propto \exp(-\frac{1}{2\sigma^2}\|\theta\|^2)$.

Posterior

$$p(\theta|X) \propto \exp\left(\sum_{i=1}^{m}\langle\phi(x_i),\theta\rangle - g(\theta) - \frac{1}{2\sigma^2}\|\theta\|^2\right)$$

This leads to the optimization problem

$$\frac{1}{m}\sum_{i=1}^{m}\phi(x_i) - \frac{1}{m\sigma^2}\theta = \partial_\theta g(\theta)$$

NATIONAL ICT AUSTRALIA

Conjugate Priors

Problem with Normal Prior

The posterior looks different from the likelihood. So many of the Maximum Likelihood optimization algorithms may not work ...

Idea

What if we had a prior which looked like additional data, that is

$$p(\theta|X) \sim p(X|\theta)$$

For exponential families this is easy. Simply set

$$p(\theta|a) \propto \exp(\langle \theta, m_0 a \rangle - m_0 g(\theta))$$

Posterior

$$p(\theta|X) \propto \exp\left((m+m_0)\left(\left\langle \frac{m\mu + m_0 a}{m+m_0}, \theta \right\rangle - g(\theta)\right)\right)$$

NATIONAL ICT AUSTRALIA