CS-567 Machine Learning

Nazar Khan

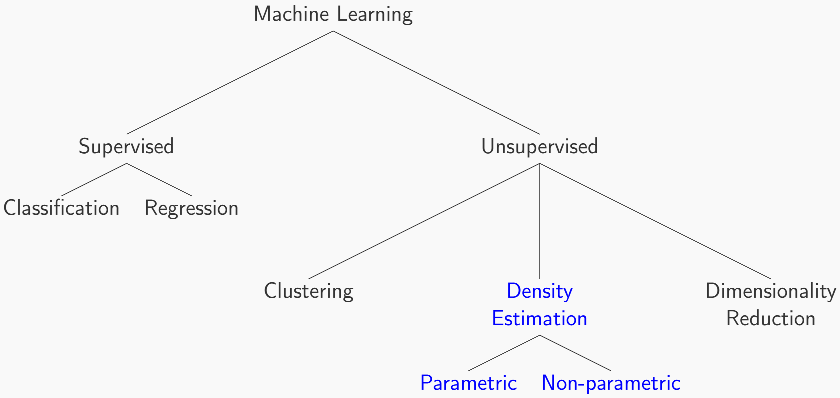
PUCIT

Lectures 14-17
Jan 12, 14, 19, 21 2015

Why study distributions?

- So that we can model unknown $p(x)$ given data $\{x\}$ corresponding to observations of random variable $\mathbf{x}$.
- Also called **density estimation**.
- Fundamentally ill-posed problem because infinitely many distributions can give rise to the obeserved data.
  - *Any* distribution that is non-zero at the observed data points *could* have generated the data.
- Chosing an appropriate distribution relates to model selection.

Density Estimation



Parametric density estimation

- A parametric distribution $p(\mathbf{x}|\boldsymbol{\theta})$ is one where parameters $\boldsymbol{\theta}$ determine the exact probability function. For example, Gaussian $\mathcal{N}(\mu, \sigma^2)$.
- Density estimation $\implies$ finding $\boldsymbol{\theta}^*$ given observed data.
  - *Frequentist approach*: Maximise likelihood $p(\text{data}|\boldsymbol{\theta})$.
  - *Bayesian approach*: Use prior $p(\boldsymbol{\theta})$ to obtain posterior $p(\boldsymbol{\theta}|\text{data})$ via Bayes' theorem and maximise it.

Non-parametric density estimation

- One weakness of parametric methods is that the functional form of the density is fixed and can be inappropriate for a particular application.
  - For example, assuming Gaussian when the observed data is not normally distributed at all (multi-modal).
- We will consider 3 non-parametric methods
  - Histograms
  - Nearest-neighbours
  - Kernels

Binary Random Variables – Bernoulli Distribution

- Can take only 2 states. That is $x \in \{0, 1\}$.
- $p(x = 1) = \mu$ and $p(x = 0) = 1 - \mu$ where parameter μ can be interpreted as the probability of success.
- Note that we can write $p(x) = \mu^x(1 - \mu)^{1-x}$. This is also called the **Bernoulli distribution**

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

Verify that this probability distribution

- is normalised,
- $\mathbb{E}[x] = \mu$, and
- $\text{var}[x] = \mu(1 - \mu)$

Bernoulli Distribution

- Likelihood for i.i.d Bernoulli data $\mathcal{D}$ is $p(\mathcal{D}|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$.
- Log-likelihood is

$$\begin{aligned}\ln p(\mathcal{D}|\mu) &= \sum_{n=1}^{N} x_n \ln \mu + (1-x_n)\ln(1-\mu) \\ &= \ln \mu \sum x_n - \ln(1-\mu)\sum x_n + N\ln(1-\mu)\end{aligned}$$

- Note that log-likelihood depends on data *only through the sum* $\sum x_n$. So $\sum x_n$ is a **sufficient statistic** for the the data under this distribution.
  - Knowing the sum is sufficient for computing the log-likelihood. The individual data points are not required.

Bernoulli Distribution

- Setting the derivative of the log-likelihood w.r.t μ to zero, we see that $\mu_{ML} = \frac{1}{N}\sum x_n = \frac{m}{N}$ where m is the number of successes (x=1) in the observed data.
- So μ_{ML} is the fraction of successes (x=1) in the observed data.
- Biased towards the observed sample (over-fitting). Solution: Use prior on μ (Bayesian approach).

Binomial Distribution

- A **binomial random variable** x measures the *number of successes in N trials*.

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

where $\binom{N}{m} = \frac{N!}{(N-m)!m!}$ is the number of ways of choosing m items from a total of N items. Explain why.

- $\mathbb{E}[m] = N\mu$. Prove it.
- $\text{var}[m] = N\mu(1-\mu)$. Prove it.

Sequential Learning

- Since posterior $\propto$ likelihood $\times$ prior, if prior has the same *functional* form as the likelihood, the posterior will also have the same functional form.
  - Gaussian likelihood $\times$ Gaussian prior leads to Gaussian posterior.
- **Now this posterior p(model|data) can be used as a prior p(model) for subsequent data.**
- This is called **sequential learning**.
- Such a prior is called a **conjugate prior**.

Sequential Learning
Beta Distribution

- Functional form of likelihood for i.i.d Binomial data is $\mu^x(1-\mu)^{1-x}$.
- A prior of the same functional form is given by the so-called **Beta distribution**

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

  where $\Gamma(x) = \int_0^x u^{x-1}e^{-u}du$ is called the gamma function.
- a and b are *hyperparameters* since the control the distribution of parameter μ.
- Verify that the beta distribution is
  - is normalised $\int_0^1 \text{Beta}(\mu|a,b)d\mu = 1$,
  - $\mathbb{E}[\mu] = \frac{a}{a+b}$, and
  - $\text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$.

Sequential Learning

Putting it all together

- Likelihood for i.i.d Binomial data is

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

- Conjugate prior is given by the beta distribution

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

- After multiplying likelihood and prior, the posterior can be written in the form

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

which is again a beta distribution.

Sequential Learning

Putting it all together

- So we can find the normalizing coefficent too and the posterior becomes

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

- Compared to prior, posterior increases a by m and b by l.
- So hyperparameters a and b can be interpreted as effective successes and failures.
- *For subsequent data*, we can treat posterior as prior and keep updating it.
  - Multiply $\underbrace{\text{current posterior}}_{\text{prior}}$ by the likelihood of the new observation. For beta distribution, increment a by 1 for $x = 1$ and b by 1 for $x = 0$.
  - Normalize.

Sequential Learning

- Sequential learning is useful for
  - online (real-time) learning because observations can be used in small batches (or one at a time).
  - large data sets because observations can be discarded after using.
- Sequential learning requires
  1. i.i.d data so that likelihood for new observation can be multiplied by the old likelihood.
  2. conjugate prior so that posterior does not change form and can be continuously updated.

Multinomial Random Variables

- Random variables that can take 1-of-K values.

$$\mathbf{x} = (0, 0, 1, 0, 0, 0)^T$$

represents an observation of $\mathbf{x}$ in which $x_3 = 1$.

- Note that $\sum_{k=1}^{K} x_k = 1$.
- If $p(x_k = 1) = \mu_k$, then $\mu_k \geq 0$, $\sum_{k=1}^{K} \mu_k = 1$ and $p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$.
- A generalization of the binomial distribution is the **multinomial distribution**

$$\mathrm{Mult}(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

where m_k is the number of data points having the k_{th} value.

Multinomial Random Variables

Sequential Learning

- The corresponding conjugate prior is given by the **Dirichlet distribution**

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

- Multiplying the multinomial likelihood with the Dirichlet conjugate prior gives a Dirichlet posterior $\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha} + \mathbf{m})$.
- This allows sequential learning for multinomial random variables.