Fundamentals of bayesian statistics

Course of Machine Learning
Master Degree in Computer Science
University of Rome ``Tor Vergata''

Giorgio Gambosi

a.a. 2018-2019

Classical (frequentist) statistics

- Interpretation of probability as frequence of an event over a sufficiently long sequence of reproducible experiments.
- Parameters seen as constants to determine

Bayesian statistics

- Interpretation of probability as degree of belief that an event may occur.
- Parameters seen as random variables

Bayes' rule

Cornerstone of bayesian statistics is Bayes' rule

$$p(X = x|\Theta = \theta) = \frac{p(\Theta = \theta|X = x)p(X = x)}{p(\Theta = \theta)}$$

Given two random variables X, Θ, it relates the conditional probabilities $p(X = x|\Theta = \theta)$ and $p(\Theta = \theta|X = x)$.

Given an observed dataset $\mathbf{X}$ and a family of probability distributions $p(x|\Theta)$ with parameter Θ (a probabilistic model), we wish to find the parameter value which best allows to describe $\mathbf{X}$ through the model.

In the bayesian framework, we deal with the distribution probability $p(\Theta)$ of the parameter Θ considered here as a random variable. Bayes' rule states that

$$p(\Theta|\mathbf{X}) = \frac{p(\mathbf{X}|\Theta)p(\Theta)}{p(\mathbf{X})}$$

Interpretation

- $p(\Theta)$ stands as the knowledge available about Θ before $\mathbf{X}$ is observed (a.k.a. prior distribution)
- $p(\Theta|\mathbf{X})$ stands as the knowledge available about Θ after $\mathbf{X}$ is observed (a.k.a. posterior distribution)
- $p(\mathbf{X}|\Theta)$ measures how much the observed data are coherent to the model, assuming a certain value Θ of the parameter (a.k.a. likelihood)
- $p(\mathbf{X}) = \sum_{\Theta'} p(\mathbf{X}|\Theta')p(\Theta')$ is the probability that $\mathbf{X}$ is observed, considered as a mean w.r.t. all possible values of Θ (a.k.a. evidence)

Definition

Given a likelihood function $p(y|x)$, a (prior) distribution $p(x)$ is conjugate to $p(y|x)$ if the posterior distribution $p(x|y)$ is of the same type as $p(x)$.

Consequence

If we look at $p(x)$ as our knowledge of the random variable x before knowing y and with $p(x|y)$ our knowledge once y is known, the new knowledge can be expressed as the old one.

The Beta distribution is conjugate to the Bernoulli distribution. In fact, given $x \in [0,1]$ and $y \in \{0,1\}$, if

$$p(\phi|\alpha,\beta) = \text{Beta}(\phi|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\phi^{\alpha-1}(1-\phi)^{\beta-1}$$
$$p(x|\phi) = \phi^x(1-\phi)^{1-x}$$

then

$$p(\phi|x) = \frac{1}{Z}\phi^{\alpha-1}(1-\phi)^{\beta-1}\phi^x(1-\phi)^{1-x} = \text{Beta}(x|\alpha+x-1,\beta-x)$$

where Z is the normalization coefficient

$$Z = \int_0^1 \phi^{\alpha+x-1}(1-\phi)^{\beta-x}d\phi = \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+x)\Gamma(\beta-x+1)}$$

Examples of conjugate distributions: beta-binomial

The Beta distribution is also conjugate to the Binomial distribution. In fact, given $x \in [0,1]$ and $y \in \{0,1\}$, if

$$p(\phi|\alpha,\beta) = \text{Beta}(\phi|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\phi^{\alpha-1}(1-\phi)^{\beta-1}$$

$$p(k|\phi,N) = \binom{N}{k}\phi^{k}(1-\phi)^{N-k} = \frac{N!}{(N-k)!k!}\phi^{N}(1-\phi)^{N-k}$$

then

$$p(\phi|k,N,\alpha,\beta) = \frac{1}{Z}\phi^{\alpha-1}(1-\phi)^{\beta-1}\phi^{k}(1-\phi)^{N-k} = \text{Beta}(\phi|\alpha+k-1,\beta+N-k-1$$

with the normalization coefficient

$$Z = \int_0^1 \phi^{\alpha+k-1}(1-\phi)^{\beta+N-k-1}d\phi = \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+k)\Gamma(\beta+N-k)}$$

Examples of conjugate distributions: dirichlet-multinomial

Assume $\boldsymbol{\phi} \sim \text{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha})$ and $z \sim \text{Mult}(z|\boldsymbol{\phi})$. Then,

$$
\begin{aligned}
p(\boldsymbol{\phi}|z,\boldsymbol{\alpha}) &= \frac{p(z|\boldsymbol{\phi})p(\boldsymbol{\phi}|\boldsymbol{\alpha})}{p(z|\boldsymbol{\alpha})} = \frac{\phi_z p(\boldsymbol{\phi}|\boldsymbol{\alpha})}{\int_{\boldsymbol{\phi}} p(z|\boldsymbol{\phi})p(\boldsymbol{\phi}|\boldsymbol{\alpha})d\boldsymbol{\phi}} \\
&= \frac{\phi_z p(\boldsymbol{\phi}|\boldsymbol{\alpha})}{\int_{\boldsymbol{\phi}} \phi_z p(\boldsymbol{\phi}|\boldsymbol{\alpha})d\boldsymbol{\phi}} = \frac{\phi_z p(\boldsymbol{\phi}|\boldsymbol{\alpha})}{E[\phi_z|\boldsymbol{\alpha}]} \\
&= \frac{\alpha_0}{\alpha_z}\frac{\Gamma(\alpha_0)}{\prod_{j=1}^K \Gamma(\alpha_j)}\phi_z \prod_{j=1}^K \phi_j^{\alpha_j - 1} \\
&= \frac{\Gamma(\alpha_0 + 1)}{\prod_{j=1}^K \Gamma(\alpha_j + \delta(j=z))} \prod_{j=1}^K \phi_j^{\alpha_j + \delta(j=z) - 1} = \text{Dir}(\boldsymbol{\phi}|\boldsymbol{\alpha}')
\end{aligned}
$$

where $\boldsymbol{\alpha}' = (\alpha_1, \ldots, \alpha_z + 1, \ldots, \alpha_K)$