Statistical Data Analysis 2

Mateusz Kapusta

6 października 2022

1 Start

Grading rules

- 50% exam (test as usuall)
- 15% and 15% for two labolatory projects
- 15% midterm test
- 5% lab activity

Pass with 50% as usual.

2 Lets gooooo

First some formulas. We denote joint probability of $mathcalD$ and Y $P(X, Y)$. As we know:

$$P(\mathcal{D}) = \sum_{Y} P(X, Y) \tag{1}$$

. Then we have conditional probability

$$P(\mathcal{D}|Y) = \frac{P(X, Y)}{P(Y)} \tag{2}$$

Theeeen

$$P(\mathcal{D}|Y) = P(Y|X)\frac{P(X)}{P(Y)} \tag{3}$$

aka Bayes theorem.

2.1 Statistical inference

Let try to reanalyse coin toss experiment. Let's assume that we have θ as our propability of heads. When we act in frequentist approach we want to estimate parameter θ using methodes as MLE. What we need to claryify we belive there is God's rule that there exist one and only θ value. In Bayesian framework we do not think aobut tru parameter but rather conditional probability $P(\theta|mathcalD)$ (X is observed data). Lets introduce

$$L(\theta) = P(mathcalD|\theta) \tag{4}$$

As we know

$$P(\mathcal{D}|\theta) = \theta^k(1-\theta)^{N-k} \tag{5}$$

When we have N tosses and k sucesses. We can of course use ML estimator and will in the limit converge. We take logliklihood

$$l(\theta) = k\log\theta + (N-k)\log(1-\theta) \tag{6}$$

In ML approach we know that $\hat{\theta} = \frac{k}{N}$ but we want more then point estimator! In frequentist approach we know we can use something like repeat data generating process but we can clearly see that this approach isn't going to be very usefull. There is better way, we can use propability to obtain k times sucess.

$$P(\hat{\theta}) = \theta^{\hat{\theta}N}(1-\theta)^{N(1-\hat{\theta})}\binom{N}{\hat{\theta}N} \tag{7}$$

So we know that we can in fact obtain prob denisty for $\hat{\theta}$. It is possible to conduct analysis of propability when we do something like bootstrap. In bayesian statistics we think about prior. It is purly our belif about the propability of parameter. Lets decompose 3.

- $P(\theta|\mathcal{D})$ - posterior
- $P(\mathcal{D}|\theta)$ - likelihood
- $P(\theta)$ - prior
- $P(\mathcal{D})$ - constant

Sometimes life is hard (and we need to use MCMC), sometimes is easy (if we choose easy prior known and conjugate prior) so choose wisely. One of the conjugate priors is beta distribution (for binomial likelihood).

$$\mathcal{B}(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{8}$$

This distribution is defined on $[0,1]$ and if $\beta = 1$ and $\alpha = 1$ we get uniform prior (in other cases it looks very diffrent so we can choose something that suits us). We also have mean of the distribution at $\mu = \frac{\alpha}{\beta+\alpha}$ so making $\alpha >> \beta$ makes distribution shifted to the right. Lets rewind

$$P(\mathcal{D}|\theta) = \binom{N}{k}\theta^k(1-\theta)^{N-k} \tag{9}$$

When we use Beta distribution as prior we get posterior

$$P(\theta|\mathcal{D}) = \mathcal{B}(\theta|k+\alpha, N-k+\beta) \tag{10}$$

so we know that posterior is very nice. Now lets introduce some new point estimators

- MAP (Maximum aposteriori estimate) $\theta_{MAP} = argmax_\theta P(\theta|\mathcal{D})$
- ML (Maximum likelihood estimate) $\theta_{ML} = argmax_\theta P(\mathcal{D}|\theta)$

In the limit of the number of samples both estimators converge to the same value due to the fact that prior is dominativ. With know data prior dominates and things differ.

3 Bayesian networks

Bayesian network consist of

- directed acyclic graph (DAG) $G = (V, E)$
- local propability distribution, one for each vertex

Propabilit distribution for joint values $X = (X_1, \ldots, X_l)$ is

$$P(X) = \Pi_i P(X_i|pa(X_i)) \tag{11}$$

so we just multiply over conditional propabilities between node and it's parrent. We can consider something like linear gaussian model so we have

$$P(X_n|pa(X_n)) = Norm(b_n + \omega_n^t X_{pa(n)}, va_n) \tag{12}$$

so each vertex is charaterized by two values, mean and covariance matrix.