

Statistical Data Analysis 2

Mateusz Kapusta

12 stycznia 2023

1 Start

Grading rules

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

Pass with 50% as usual.

2 Lets gooooo

First some formulas. We denote joint probability of \mathcal{D} and Y $P(X, Y)$. As we know:

$$P(\mathcal{D}) = \sum_Y P(X, Y) \quad (1)$$

. Then we have conditional probability

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

Theeeen

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

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 $P(\theta|\mathcal{D})$ (X is observed data). Lets introduce

$$L(\theta) = P(\mathcal{D}|\theta) \quad (4)$$

As we know

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

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

$$l(\theta) = k \log \theta + (N - k) \log (1 - \theta) \quad (6)$$

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.

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

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).

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

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

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

When we use Beta distribution as prior we get posterior

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

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

- MAP (Maximum a posteriori estimate) $\theta_{MAP} = \operatorname{argmax}_{\theta} P(\theta|\mathcal{D})$
- ML (Maximum likelihood estimate) $\theta_{ML} = \operatorname{argmax}_{\theta} P(\mathcal{D}|\theta)$

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

3 Bayesian networks

Bayesian network consist of

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

Probability distribution for joint values $X = (X_1, \dots, X_l)$ is

$$P(X) = \prod_i P(X_i | pa(X_i)) \quad (11)$$

so we just multiply over conditional probabilities between node and its parent. We can consider something like linear gaussian model so we have

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

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

3.1 Markov Blanket

Markov blanket is subset of Bayesian is set of parents, coparents and children for given vertex. It is useful as it contains all informations to calculate conditional probability of value of given vertex. We can say, that $P(V|all) = P(V|MB(V))$. Simplifying things we can say that

$$P(V|all) = P(V|MB(V)) = \frac{\sum P(V|parents) + P(children|V, coparents)}{Normalization} \quad (13)$$

So it is important when we are sampling using MCMC.

3.2 Conditional independence

We say that A and B are conditional independent given C and write $A \perp B|C$. If A and B are children of C then we have conditional independence, the same when there is connection between A and B through C but not when C is child of A and B . In fact, when A and B meet head to head we have independence but not conditional independence. Now let write some more definitions

- let A and B and z is set of other nodes
- If the arrows between our points meet head to head and head to tail path is blocked.
- If the arrows meet at node C head to head and neither the node nor its descendance belong to z (no "explaining" away)

4 Time to learn

We can decompose learning process to parameter learning (modeling parameters of graph) and structure learning (determining structure of network). We can know structure of graph, know all data or any

combination.

4.1 Known graph, known data

Well, we can just use MAP estimation and we are fine. We treat given parameter as random variable (hidden), assume we have some hyperparameters, we take some prior and then just use MAP. In discrete case we use something like binomial distribution. In case of many possible outcomes we can use multinomial distribution (with conjugate prior in form of dirichlet distribution).

4.2 Known data, unknown graph

Well, just want to know the model, we can write

$$P(G|D) \propto P(D|G)P(G) = \int P(D|G, \theta)P(\theta|G)P(G)d\theta \quad (14)$$

It's bad but we can sample. One important thing - Bayes factor $\frac{p(\mathcal{D}|M_i)}{p(\mathcal{D}|M_j)}$. It is used to decide which model is better.

4.3 Markov chains as examples of bayesian networks

Lets think of the chain that is composed of N parts X_i . Because we have chain X_i is dependent only on X_{i-1} . When we have Markov property

$$P(X_i|X_{i-1}) = P(X_2|X_1) \quad (15)$$

. As always when dealing with markov chains we have also classical transition matrix T . Example: CpG islands. We consider DNA sequence. Sometimes CG pairs will be more often than others what is called CpG island. When we have supervised learning and somebody gave us some data labeled as cpg or not cpg islands we can imagine, that we have two models for two cases labeled T^+ , T^- . We then can use something like log-odds

$$S(X) = \log \frac{P(X|T^+)}{P(X|T^-)} \quad (16)$$

. We can set a threshold for classification and we have a classifier!

5 (Gaussian) Mixture Models

Lets get into, we have data, that is clustered (n components, d dimensional data). Each component have weights π_i and are assumed to represent probability of sampling the object from given subtype. We have hidden random variable z_i and we assume

$$\begin{aligned} Z_j &\sim \text{Mult}(\pi, 1) \\ x_i|z_i &\sim \mathcal{N}(\mu_z, \Sigma_z) \end{aligned}$$

We do not observe to which cluster observation belongs! In case of clustering we just want to infer what is z_i , we can do it using what we previously introduced so

$$P(z_i = j|x_i) \propto \pi_j \phi(x_i|\mu_j, \Sigma_j) \quad (17)$$

We can also imagine, that we do not know π_j or parameters of gauss distributions. In this case we can maximize log-likelihood of our data with respect to parameters. Well, in case when we have only one distribution it is easy, if not we are essentially screwed as there is no answer in closed form. We need to know z_i to estimate parameters if we want to work in case $n > 1$ as we can introduce something like complete log-likelihood (we know not only x_i but also z_i). But there is also one more way...

5.1 Expectation Maximization

This is what we can do

- Initialize our variables, z_i and μ_j, Σ_j
- Alternate until convergence
 - Compute soft class memberships given current parameters (soft means distribution of probabilities, not point estimate)
 - Use point estimates to determine cluster membership, estimate π_i and then compute μ_j, Σ_j using techniques we previously mentioned.

What we see, that we take some values, perform easy task that we described previously. We always approximate and then iterate. It can be shown that this algorithm converges but only to local minimum.

Time for some cool calculations, there is hidden variable Z , observed variables X and parameters of Z and X λ and θ respectively:

$$l_{obs}(\theta, \lambda) = \log \sum_Z P(X, Z|\theta, \lambda) \quad (18)$$

$$= \log \sum_Z q(z) \frac{P(X, Z|\theta, \lambda)}{q(z)} \quad (19)$$

$$= \log \quad (20)$$

5.2 Kullback-Leibler divergence

The KL divergence measures distance between the probability distributions $Q(X), P(X)$.

$$D_{KL}(P||X) = - \sum_X P(X) \log \frac{Q(X)}{P(X)} \quad (21)$$

. This is always nonnegative and equal to 0 only, if $Q(X) = P(X)$. what is important this is not symmetrical! After some math we obtain, that

$$\mathcal{L}(\theta) = F(q, \theta) + D_{KL}(q||P) \quad (22)$$

.

6 Variational inference

What we want to do? Basically we want to approximate $P(Z|X) \approx Q(Z)$. Now let's go to ELBO

6.1 ELBO

let's use $P(Z|X)$ and plug it into KL divergence

$$G_{KL}(Q||P) = \sum_Z Q(Z) \left[\log \frac{Q(Z)}{P(Z.X)} + \log P(X) \right] \quad (23)$$

$$= \mathcal{E}_Q [\log Q(Z) - \log P(Z.X)] + \log P(x) \quad (24)$$

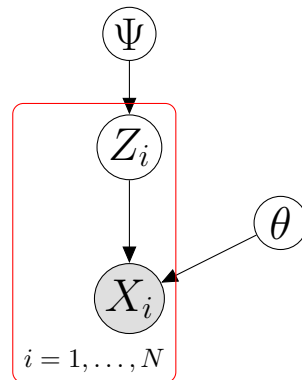
Hence

$$\log P(X) = D_{KL}(P||Q) + F(Q) \quad (25)$$

where $F(Q)$ is called evidence lower bound (ELBO). In EM case we were taking substituting $Q(Z) = P(Z|X)$ because we had analitical forms of $P(Z|X)$, this time we need to make more approximations. How to do it? We can take something like general set of functions and just to use some software for nonlinear optimization. There is also other way called method of factorized distributions. We assume, there is factorization of Z into separate domasins Z_i so it can be factorized as joint propability of few distributions.

7 Variational Autoencoders

This time we will cover theorey behind Variational Autoencoders.



Rysunek 1: Bayesian net model for variational autoencoder

In autoencoders we have some prior distribution Ψ . In the case of variational autoencoder we are modeling $P(X_i|Z_I)$ by neural network making it very powerfull model. Variational autoencoders are generaiv models so they allow us to generate new data given a sample model. $p_\theta(x|z)$ is posterior modelled by neural network called decoder. p_ψ is reffered as prior distribution. Our model can be we are modeling:

- Soft propability
- Propability of

. We can also write down ELBO for VAE:

$$ELBO = \mathbb{E}_{q_\theta}[\log p_\theta(X|Z)] - D_{KL}(q_\phi||p_\psi) \quad (26)$$

. This equation is nondifferentiable in our case as numerically we want to sample from the distribution and this is non differentiable. We can deal with this problem assigning

$$Z = \mu(X) + \epsilon\sigma(X) \quad (27)$$

where σ and μ are modelled by neural networks based on X and ϵ is sampled from standard normal gaussian. This is differentiable and can allow us to use backpropagation to train encoder neural network.

8 Sampling

TODO

9 Model selection

We can divide model selection in two categories, structure learning and parameter learning. Now let's introduce new measure called mutual information

$$I(X, Y) = H(X) - H(X, I) = D_{KL}(P(X, Y) || P(X)P(Y))$$

If two random variables are independent mutual information is zero! Now let's consider we have some bayesian tree T . We can measure weight of tree as

$$w(T) = \sum_i I(X_i, X_{pa(i)})$$

. We can prove that this type of selection minimalizes divergence between original joint distribution and joint distribution modelled by tree. To search for such tree we can use kruskal algorithm for maximum spanning tree with weights but weights are mutual information between variables. To compute those values one can use maximum likelihood estimator $f_{ij}(u, v) = P(u = i, v = j)$ to compute

$$\hat{I}(x_i, x_j) = \sum_{u,v} f_{ij}(u, v) \log \frac{f_{ij}(u, v)}{f_i(u)f_j(v)}$$

. Now lets try to compute Bayes factor to compare models. We want to find $P(D|G)$ so marginalized out propability of data fiven model. We know

$$P(D|G) = \int P(D|G, \theta) P(\theta|G) d\theta$$

which is hard to compute. When we assume posterior of parameter is unimodal, peaked and had a rather uninformative prior we can approximate this as

$$\log P(D|G) = \log P(D|g, \hat{\theta}) + \log \frac{\Delta\theta_{posterior}}{\Delta\theta_{prior}}$$

where $\hat{\theta}$ is MAP estimator. We see we have penalization for overfitting models in the form of deviation from MAP estimate which explodes if posterior is very narrow. We can also introduce other criterions. Using Laplace expansion (assuming gaussian distribution) we can introduce

$$\log P(D|G) = \log P(D|\hat{\theta}, G) - \frac{1}{2} \log \det\{H\} + \frac{\nu}{2} \log 2\pi c^2$$

Where H is hessian matrix of evidence defined as $E(\theta) = -\log P(D|\theta, G)$. Then laplace aproximation yields

$$E(\theta) \approx E(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T \mathbf{H}(\theta - \hat{\theta})$$

. When we further assume equall curvature along direction $\epsilon_i = 2\pi N c^2$ where ϵ_i is eigenfunction of H we get

$$\log P(D|G) \approx \log(D|\hat{\theta}, G) - \frac{\nu}{2} \log N$$

where ν is dimension of parameter space. We can also want to use search and score approach, we introduce graph and then by small greedy modifications we want to find better structure. During this approach we want to score small modifications instead of score whole model. This is possible of BIC due to the fact that BIC also decomposes as likelihood and we can introduce local score instead of global one.

9.1 Structural EM