week 8: Bayesian models

wbg231

December 2022

1 frequentest stats

- parametric family is a set such that

$$p(y|\theta) : \theta \in \Theta$$

  where y is our sample space, θ is our parameter and Θ is our parameter space

- in this model we assume that there is a true data generating process given by $P(y|\theta)$ for some $\theta \in \Theta$
- if we knew the correct θ there would be know need to statics
- however we are only able to view an iid smaple of our data from the data generating process $P(y|\theta)$
- a statistic $s = s(\mathcal{D})$ is any function of our data
- a statistic $\hat{\theta} = \hat{\theta}(\mathcal{D}) : \theta \in \Theta$ is a point estimator of θ that is an estimate of some parameter that we are using for estimation
- ideal point estimates are unbiased, consistent and efficient (ie could not get a better point estimate from the data)
- MLE is consistent and efficient under some assumptions

coin flip

- our parametric family are Bernoulli random variables with parameter $\theta \in (0, 1)$
- MLE on this is fairly straight forward

$$L_{\mathcal{D}}(\theta) = P(y|x, \theta) = \theta^{n+h}(1-\theta)^{n_y}$$

- then we can max the log likelihood

$$\theta_{mle} = argmax_{\theta}(n_h log(\theta) + n_t log(1-\theta))$$

- $$\nabla_\theta = \frac{nh}{\theta} - \frac{n_t}{1-\theta} \Rightarrow \theta_{mle} = \frac{n_h}{n_h + n_t}$$
- so as we would expect the mle estimate for coin flipping is the % of tails in the dataset

Bayesian stats

- bayesian models a parametric family family of distribution to model the $y|x, \theta$ as well as a prior distribution $P(\theta)$
- putting the pieces together we get the joint density on θ and $\mathcal{D}$
$$P(\mathcal{D}, \theta) = P(\mathcal{D}|\theta)P(\theta)$$
- the posterior for *theta* is $P(\theta|\mathcal{D})$ the posterior is how we rationally update our beliefs
- note that $P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} \propto P(\mathcal{D})P(\theta)$ so that is your posterior is proprtional to the product of the likckyhood of the data and the prior
- so for the purposes of learning we are trying to find $\theta = argmax_\theta P(\theta|\mathcal{D})$ and we do this by looking at our likelihood function and prior
- so keep in mind that we are reasoning about not just the probaility of our outcomes y, but also the likelyhood of our parameter

coin flipping example

- lets model a bayesian coin flip
- we can chose our parametric family to be $P(y|\theta) = \theta$ ie a bernoulli
- beta prior a prior is beta distributed if
$$\theta \sim Beta(\alpha, \beta)$$
$$P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$
- the shape of these can vary a lot
- proving it is pretty involved by $E[\theta] = \frac{\alpha}{\alpha+\beta}$
- the mode of beta is $argmax_\theta P(\theta) = \frac{\alpha-1}{\alpha+\beta-2}$
- so with this we can calculate our posterior $P(\theta|\mathcal{D}) \propto P(\theta)P(\mathcal{D}|\theta) \propto \theta^{h-1}(1-\theta)^{t-1} \times \theta^{nh}(1-\theta)^{nt} = \theta^{h-1+n_h}(1-\theta)^{t-1+n_t}$
- note here that our posterior is beta distributed
- think of this as our prior being intial examples that are as we would expect them to be and then we start viewing furhter examples

conjugate prior

- let π be a family of prior distributions on Θ
- let P be a parametric family of distributions with parameter space Θ
- a family of distributions π is conjugate to parametric model P if for any prior in π the posterior is always in π
- the beta family is conjugate to the bernoulli family of parametric distributions
- the real point of doing bayesian methods is that at the end we have a distribution of parameters $P(\theta|D)$ on which to evaluate $P(D|\theta)$ instead of just a single estimate as we get with frequentest models

bayesian point estimates

- we have the posterior distributions $\theta|\mathcal{D}$
- how do we get a point estimate $\hat{\theta}$ from this?
- that is a choice common options are a posterior mean $\hat{\theta} = E[\theta|\mathcal{D}]$
- or Max a posterior estimate $\hat{\theta} = argmax_\theta P(\theta|\mathcal{D})$ this is the mode of the posterior

what else can we do with a posterior

- we can quantity uncertainty about our estimate
- get a 95% credible set
- select a point estimate using bayesian decision theory, where we chose a loss function adn then minimize expected risk with respect to the posterior

Bayesian decision there

intro

- we need the following ingredients
  1. a parameter space Θ
  2. a prior distributions $P(\theta) \in \Theta$
  3. an action space A
  4. a loss function $\ell : A \times \Theta \to \mathbb{R}$

- we define the posterior risk for an action $a \in A$ as

$$r(a) = E[\ell(\theta, a)|\mathcal{D}] = \int \ell(\theta, a) P(\theta|D) d\theta$$

  so this is a weighted averge loss of our action over all values of $\theta \in \Theta$

- this is a lot more robust than just choiring a single theta
- a bayes action a^* is an action that minimizes posterior risk

$$r(a^\theta) = min_{a \in A} r(a)$$

  ie the best possible action in terms of exited loss under the posterior

- squared loss $\ell(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2 \rightarrow$ (PICK the value that is clostest to all others in ℓ_2 space ie the mean)
- zero one loss $\ell(\theta, \hat{\theta}) = \mathbb{I}(\theta \neq \hat{\theta})$ posterior mode (pick the value that equals the most other values ie the mode)
- absolute loss $\ell(\hat{\theta}, \theta) = |\theta - \hat{\theta}|$ posterior medan

recap and interpretation

- the prior represents beliefs about θ before observing the data
- the posterior represents how we rationally update these bellies after seeing θ
- all interfaces and actions are based on the posterior distribution

bayesian conditional models

- we need an input and outcome space (X,Y)
- a well as a parametric family of distributions to base our likelihood function off of

$$\{P(y|x, \theta) : \theta \in \Theta\}$$

- as well as a prior $P(\theta)$
- what point estimate we use depends on our prior
- so in modeling our goal is to find a functor that takes X and produces **a distribution** on our output space
- in the frequentest approach

1. we chose a family conditional probability density
2. select one estimate from among them using mle

- in bayesian methods
  1. we chose a parametric family of conditional densities
  $$\{P(y|x,\theta) : \theta \in \Theta\}$$
  2. as we as a prior distribution $P(\theta)$
- in this context we do not need to make a discrete prediction of θ from our hypothsis space we can maintain uncertainty

prior predictive distribution

- suppose we have not yet seen any data
- in the bayesian setting we can still produce a prediction function using the prior predictive function
  $$x \to P(y|x) = \int P(Y|x,\theta)P(\theta)d\theta$$
  this is an averge on all conditional densities in our family weighted by the prior

bayesian vs frequentest approach

- in bayesian stats we have two distributions on Θ
- te prior distribution $P(\theta)$ as well as the posterior $P(\theta|D)$
- in the frequentest approach we chose a point estimate $\hat{\theta} \in \Theta$ and predict
  $$P(y|x,\hat{\theta}(D))$$
- in the bayesian approach we integrate over out over Θ wrt $P(\theta|\mathcal{D})$ and predict with
  $$P(y|x,\mathcal{D}) = \int P(y|x,\theta)P(\theta|\mathcal{D})d\theta$$
  so this is a distribution of our outcomes over our parameter space
- once we have a predictive distribution $P(y|x,\mathcal{D})$ it is easy to generate a single point depending on the loss function we are using

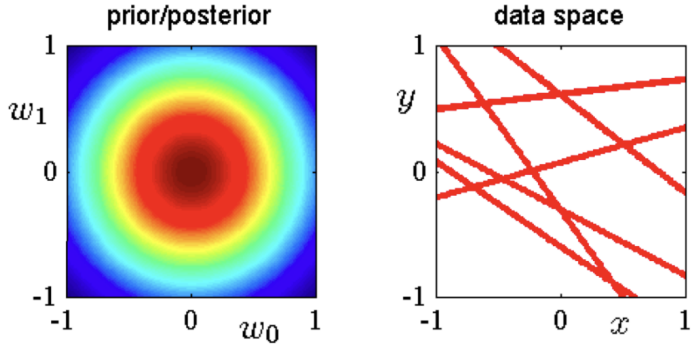
gaussian regression example

1 dimensional example

- let our input space be discrete $X = [-1, 1]$ and our output space be $y \in \mathbb{R}$
- given x $y = w_0 w_1 x + \epsilon$ where

$$\epsilon \sim \mathcal{N}(0, .2^2) \iff y|x, w_0, w_1 \sim \mathcal{N}(w_0 + w_1 x, .2^2)$$

- we know that $w \in \mathbb{R}^2 \iff \Theta = \mathbb{R}^2$ is our parameter space
- let our prior distribution be $w = (w_0, w_1) \sim (0, \frac{1}{2}I)$ (so they are a gaussian random vector centered at zero)
- our prior is a mean 0 centerd gaussian random variole with a symetric covariance matrix so teh joint pdf of the our prior has perfectly circular contour lines, if this were ture we would expect to see data liek this that follows



no real pattern

- then as we get more observations, they can over power our prior, and if tehy have a trent our pmd will shift away from teh prior

gaussian regression closed form

- to recap our model is

$$w \sim \mathcal{N}(0, \Sigma_0)$$

and

$$y_i|x, w \sim \mathcal{N}(w^t x_i, \sigma^2)$$

- our posterior is also gaussian

$$w|\mathcal{D} \sim \mathcal{N}(\mu_D, \Sigma_D)$$

- this closed form can lead to either linear or ridge regression depending on the from of our covariance matrix