Lecture 7 probabilistic models - Bayesian methods

wbg231

December 2022

1 introduction

- we have so far done frequentest probabilistic models using MLE
- we are going to use Bayesian methods to get some uncertainty around the prediction

2 classical stats

2.1 parametric family of densities

- a parametric family of densities is a set
$$\{p(y|\theta) : \theta \in \Theta\}$$
this is a set of distributions
- where $P(y|\theta)$ is a density on a sample space y, and θ is a parameter in a parameter space Θ
- this is a common starting point for Bayesian statistics.

frequentest statistics

- in classical of frequentest statistics we are working with a parametric family of distributions
$$\{P(y|\theta : \theta \in \Theta)\}$$
- however we assume that there is some $\theta_{true} \in \Theta$ which has governed the distribution of our observed data
- so if we know θ_{true} there would be no need for statistics
- but we can not view the true data generating process, as we only have a finite sample $\mathcal{D} : (y_1 ... y_n)$ generated independent and identically distributed from $P(y|\theta_{true})$

point estimation

- one type of statistical problem is point estimation
- a **statistic** $s = s(\mathcal{D})$ is any function of our data
- **a point estimator of** $\boldsymbol{\theta}$ is a function of our data $\boldsymbol{\hat{\theta} = \hat{\theta}(\mathcal{D})}$ for some $\boldsymbol{\theta \in \Theta}$
- a good point estimate will have $\boldsymbol{\hat{\theta} \approx \theta_{true}}$
- we want a point estimate to have the following properties

  1. **consistency** that is if we think of our point estimate on n data points as $\boldsymbol{\hat{\theta}_n}$ then we have $\boldsymbol{lim_{n \to \infty} \hat{\theta}_n \to \theta_{true}}$ that is as we get more data our point estimate generally more accurate, that is as we get more data our error gets lower
  2. **efficiency** formally the efficiency of an unbiased estimator $\boldsymbol{\hat{\theta}}$ of parameter $\boldsymbol{\theta}$ is $\boldsymbol{e(t) = \frac{\frac{1}{I(\theta)}}{var(\hat{\theta})}}$, where $\boldsymbol{I(\theta)}$ is a measure of the amount about how much our observed data $\boldsymbol{\mathcal{D} = \{y_1..y_n\}}$ tells us about our parameter $\boldsymbol{\theta}$ it related to entropy which is how predicable a variole is effectively. basically this is saying we want our estimator to train well on relatively little data efficency

- maximum likelihood estimators as consistent and efficient under reasonable assumptions

2.2 coin example

- we have a parametric family of mass functions $\boldsymbol{P(\text{heads}|\theta) = \theta}$ for some $\boldsymbol{\theta \in \Theta := (0, 1)}$

2.3 coin flipping mle

- suppose we have dataset $\boldsymbol{\mathcal{D} = \{H...T\}}$ whee we have $\boldsymbol{n_h}$ heads and $\boldsymbol{n_t}$ tails and the flips are iid
- the likelihood function of our data is then $\boldsymbol{\mathcal{L}_{\mathcal{D}}(\theta) = P(\mathcal{D}|\theta) = P(y_1...y_n|\theta) = \Pi_{i=1}^{n} P(y_i|\theta) = (\theta)^{n_h}(1-\theta)^{n_t}}$
- we can max the log likelihood of our data as a function of $\boldsymbol{\theta}$ as $\boldsymbol{max_{\theta \in \Theta} \ell(\theta) = max_{\theta \in \Theta} n_h log(\theta) + n_t log(1-\theta)}$
- we can then see that $\boldsymbol{\frac{\partial \ell}{\partial \theta} = \frac{n_h}{\theta} - \frac{n_t}{1-\theta}}$
- setting this equal to zero we get $\boldsymbol{\theta_{mle} = \frac{n_h}{n_h + n_t}}$ which is the empirical fraction of heads which makes sense

bayesian statistics

- in bayesian stats we introduce a prior distribution
- **the prior distribution** is defined as $\boldsymbol{P(\theta)}$, and it represents our bellies about how $\boldsymbol{\theta}$ is distributed over the parameter space $\boldsymbol{\Theta}$ prior to seeing any data

a bayesian model

- there are two prices to a bayesian model
  1. a parametric family of densities $\boldsymbol{P(\mathcal{D}|\theta \in \Theta)}$ that is basically a set of distributions we think our data given the parameter may have
  2. we alos need our prior $\boldsymbol{P(\theta) : \theta \in \Theta}$
- given both of these pieces we can write our joint density $\boldsymbol{P(\mathcal{D}, \theta) = P(\mathcal{D}|\theta)P(\theta)}$ so we have the joint density of our data and the model

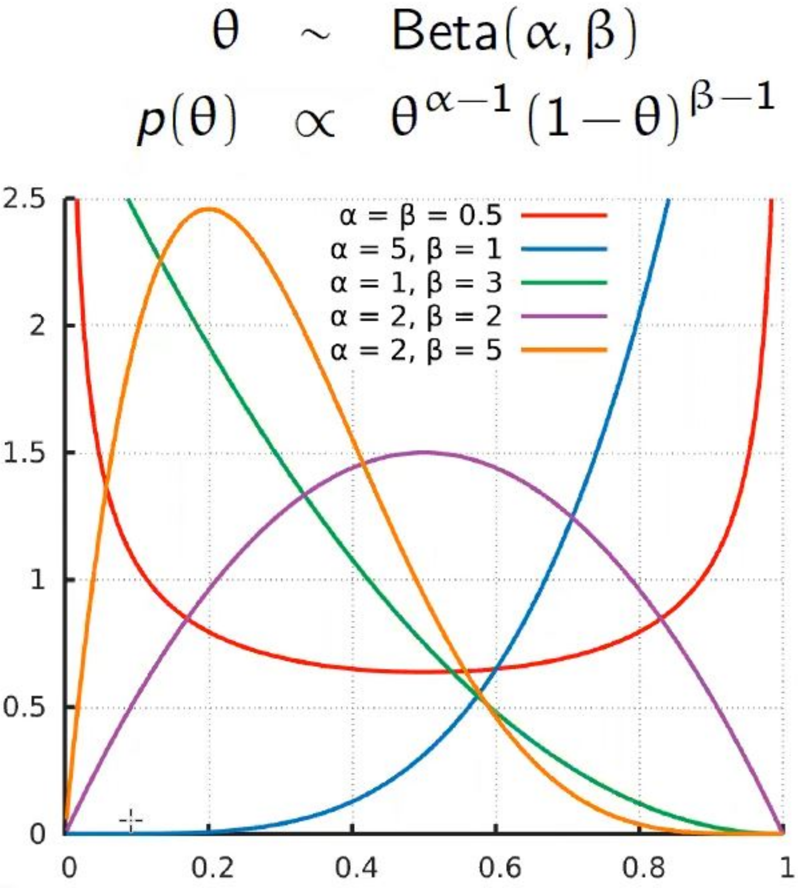
- ### the posterior

- **the posterior distribution** for $\boldsymbol{\theta}$ is $\boldsymbol{P(\mathcal{D}|\theta)}$
- the prior is our belives about the parameter before seeing any data
- the posterior represents how we rationally update our beliefs about $\boldsymbol{\theta}$ after seeing our data
- we can write the posterior as $\boldsymbol{P(\mathcal{D}|\theta) = \frac{P(\mathcal{D}),\theta)}{P(\theta)} = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}}$ where we think of both sides as a function of $\boldsymbol{\theta}$ for a fixed dataset $\boldsymbol{\mathcal{D}}$
- given our data set is fixed $\boldsymbol{P(\mathcal{D})}$ is constant so we can ignore it
- so in practice we solve for $\boldsymbol{P(\mathcal{D}|\theta)P(\theta)}$ as we know that $\boldsymbol{P(\theta|\mathcal{D}) \propto P(\mathcal{D}, \theta)P(\theta)}$

coin flipping bayesian example

- supper we have a parametric family of mass functions $\boldsymbol{P(heads|\theta) = \theta}$ where $\boldsymbol{\theta \in \Theta = (0, 1)}$
- we need a prior distribution $\boldsymbol{P(\theta)}$ on $\boldsymbol{\Theta}$
- typically we chose a distribution from the beta family

beta distributions

- given we assume our prior is $\boldsymbol{\theta} \sim \text{beta}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ we know that $\boldsymbol{P(\theta)} \propto \boldsymbol{\theta^{\alpha-1}(1-\theta)^{\beta-1}}$
- the beta family takes two parameters
- given these parameters we get a distribution over $\boldsymbol{\theta}$ but notice that this distribution is independent of our data
- the shape of the beta distribution can vary a lot.
- 

- the beta distribution is nice because it is only defined in 0,1
- given that $\boldsymbol{\theta} \sim \text{beta}(\boldsymbol{\alpha}, \boldsymbol{\beta})$
- $\boldsymbol{E}[\boldsymbol{\theta}] = \frac{\boldsymbol{\alpha}}{\boldsymbol{\alpha}+\boldsymbol{\beta}}$ the proof is kind of tedious so i am just going to link it

- then we can find the mode of the beta distribution. the mode is the most common value in a pdf that is $\boldsymbol{argmax_{\theta\in[0,1]}P(\theta) = \frac{\alpha-1}{\alpha+\beta-2}}$
- showing the mode is a big less tedious classically just note $\boldsymbol{P(\theta) = \frac{1}{B(a,b)}\theta^a(1-\theta)^b}$, then we can take the derivative and solve for $\boldsymbol{\theta^*}$

coin flipping prior

- lets say that our prior is $\boldsymbol{\theta \sim beta(c,c)}$ where $\boldsymbol{h = c = t}$ that is we are assuming that the likelihood of heads and tails are equal then our mean and mode are both equal to $\boldsymbol{\frac{1}{2}}$

coin flipping posterior

- so our likelihood of the deta given $\boldsymbol{\theta}$ is $\boldsymbol{\mathcal{L}_{\mathcal{D}}(\theta) = P(\mathcal{D}|\theta) = \theta^{n_h}(1-\theta)^{n_t}}$
- then our posterior density is

$$\boldsymbol{P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta) \propto \theta^{n_h}(1-\theta)^{n_t}\times(1-\theta)^t\theta^h = \theta^{h-1+n_h}(1-\theta)^{t-1+n_t}}$$

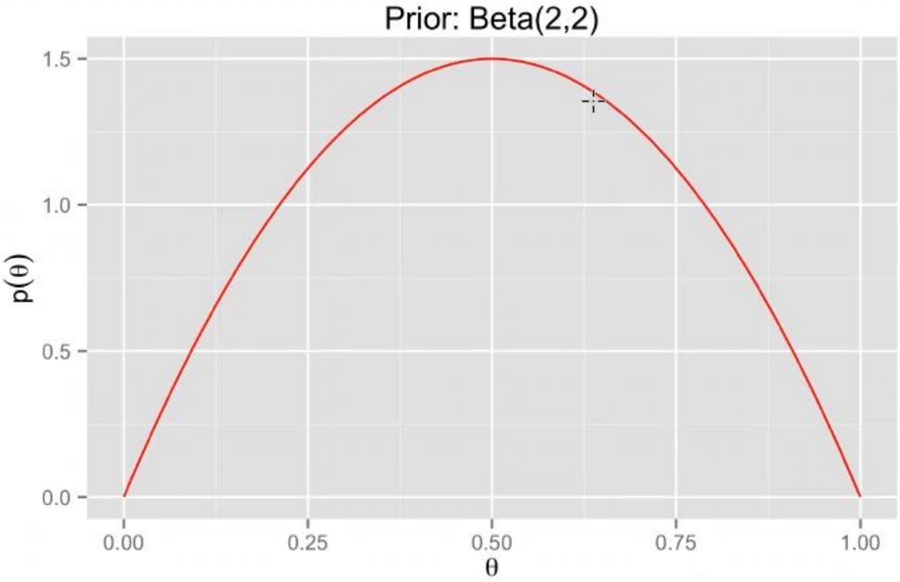
- note that our posterior is in the beta family that is $\boldsymbol{P(\theta|d) \propto \theta^{h-1+n_h}(1-\theta)^{t-1+n_t} \Rightarrow \theta|\mathcal{D} \sim beta(h+n_h, t+n_t)}$
- as the number of coin flips goes to infinity the prior will matter less as the values of h and t are fixed and $\boldsymbol{n_h, n_t}$ grow, so when we have a lot of data we weigh it quite heavily
- and when we do not have much data we rely more on our prior

conjugate priors

- in this case the posterior was in the same family of distributions as the prior
- this makes the math easy
- let $\boldsymbol{\pi}$ be a family of posterior distributions on $\boldsymbol{\Theta}$
- let $\boldsymbol{P}$ a parametric family of distributions on parameter space $\boldsymbol{\Theta}$
- family of distributions $\boldsymbol{\pi}$ is **conjugate** to the parametric model P if $\boldsymbol{\forall p(\theta) \in \pi}$ (that is all priors on $\boldsymbol{\pi}$) the posterior is in $\boldsymbol{pi}$ that is $\boldsymbol{P(\theta) \in \pi \Rightarrow P(\theta|\mathcal{D}) \in \pi}$
- the beta family is conjugate to a bernoulli model
- this is not easy to do, but it is nice when it works

coin flipping concrete example

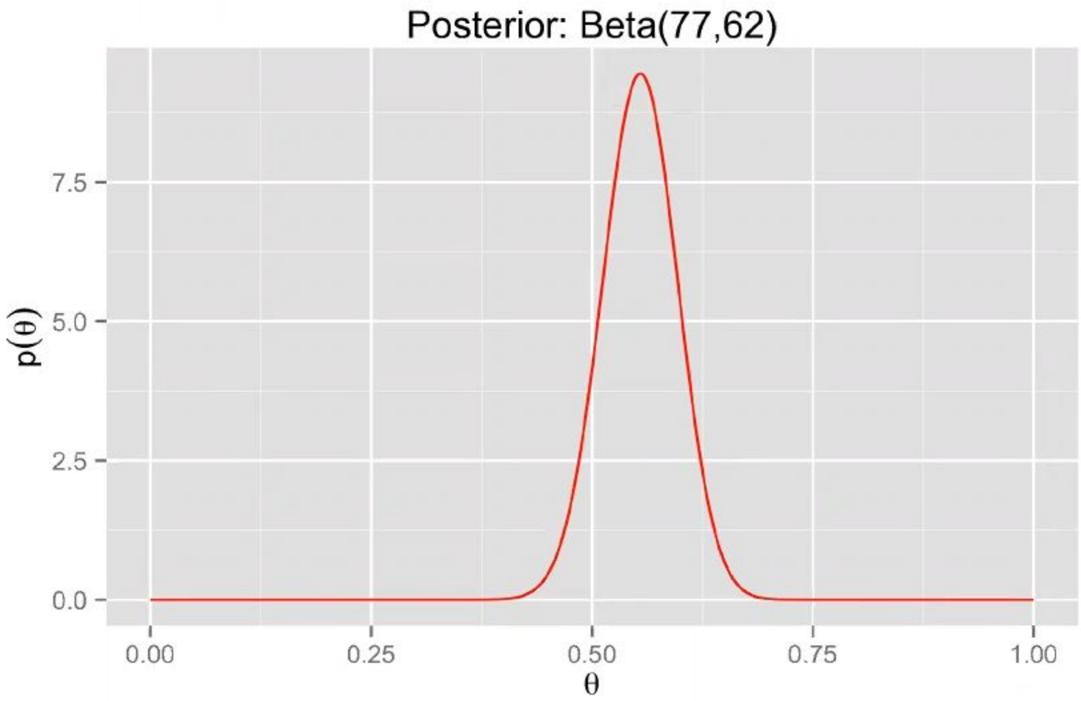
- suppose we a parametric probabilistic model of our coin $\boldsymbol{P}(\text{heads}|\boldsymbol{\theta}) = \boldsymbol{\theta}$
- and our parameter space is $\boldsymbol{\theta} \in \boldsymbol{\Theta} = [\mathbf{0}, \mathbf{1}]$
- and our prior is $\boldsymbol{\theta} \sim \boldsymbol{beta}(\mathbf{2}, \mathbf{2})$



-
- our prior assumes that $\boldsymbol{\theta}$ will be distributed kind of evenly but centered at 0

with data

- suppose we have some data, where we saw 75 heads and 60 tails
- doing maximum liklyhood estimation we would get $\hat{\boldsymbol{\theta}}_{\boldsymbol{mle}} = \mathbf{.556}$
- using bayesian methods we would get a posterior $\boldsymbol{\theta}|\boldsymbol{\mathcal{D}} \sim \boldsymbol{beta}(\mathbf{77}, \mathbf{62})$



-

- as we can see both estimates are centered at about 55
- but doing it with bayesian methods we have a distribution of parameters as opposed to just a single estimate of $\boldsymbol{\theta}$

bayesian point estimates

- what if want to give a point estimate from our posterior
- there are a few common options
  1. posterior mean $\hat{\boldsymbol{\theta}} = \boldsymbol{E}[\boldsymbol{\theta}|\boldsymbol{\mathcal{D}}]$
  2. maximum a posteriori estimate (MAP) $\hat{\boldsymbol{\theta}} = \boldsymbol{argmax_{\theta} P(\theta|\mathcal{D})}$ (this is the mode of the posterior)

what else can we do with a prior

- we can use it to quantify our uncertainty around the estimate
- we could make a credible set for $\boldsymbol{\theta}$ that is an interval $[\boldsymbol{a}, \boldsymbol{b}]$: $\boldsymbol{P(\theta} \in [\boldsymbol{a}, \boldsymbol{b}]|\boldsymbol{\mathcal{D}}) \geq \boldsymbol{\alpha}$ this is effectively bayesian confidence interval

- we could also select a point estimate using bayesian decisions theory. this requires us to chose a loss function, and chose an action to minimize the expected risk with respect to

3 bayesian decision theory

- we need the following ingredients
  1. parameter space $\boldsymbol{\Theta}$
  2. prior $\boldsymbol{p(\theta) : \theta \in \Theta}$
  3. action space A
  4. loss function $\boldsymbol{\ell : A \times \theta \Rightarrow} \mathbb{R}$
- the output is no longer the parameter it is an action
- the posterior risk of an action $\boldsymbol{a \in A}$ (also alled tthe expected loss under the posterior)is

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

- so that is more or less looking at a weighted average of our action over all possible values of theta
- this is more robust as we are only choosing an action a, not a point estimate of our parameter
- a byes action $\boldsymbol{a* = min_{a \in A} r(a)}$ the action that minimizes risk, ie the best action

bayesian point estimation

- we have a data set $\boldsymbol{\mathcal{D}}$ genrated by $\boldsymbol{P(y|\theta)}$ for some unkown $\boldsymbol{\theta \in \Theta}$
- we want a point estiamte for $\boldsymbol{\theta}$
- we need to chose a prior $\boldsymbol{P(\theta) : \theta \in \Theta}$ and loss $\boldsymbol{\ell(\hat{\theta}, \theta)}$
- and our goal is to fund the action $\boldsymbol{\hat{\theta}}$ that minimizes the posterior risk.
- the point of this is that bayesian point estimation can be looked with the bayesian decision theory framework

important cases

- squared loss $\ell(\theta - \hat{\theta}) = (\theta - \hat{\theta})^2$ minimizing this gives the $\hat{\theta} =$ posterior mean
- zero one loss $\ell(\theta - \hat{\theta}) = 1(\theta \neq \hat{\theta})$ minimizing this gives the $\hat{\theta} =$ posterior mode (not a good idea for when θ is continuous)
- $\ell(\theta - \hat{\theta}) = |\theta - \hat{\theta}|$ minimizing this gives the $\hat{\theta} =$ posterior median

squared loss example

- want to find an action $\hat{\theta} \in \Theta : \hat{\theta} \in argmin_{\theta \in \Theta} \int (\theta - \hat{\theta})^2 P(\theta|\mathcal{D})d\theta$
- we can take the partial of our risk as $\frac{\partial r(\hat{\theta})}{\partial \hat{\theta}} = -2\int(\theta - \hat{\theta})p(\theta|\mathcal{D})d\theta$ $= -2\int \theta P(\theta|\mathcal{D})d\theta + 2\theta \int P(\theta|\mathcal{D})d\theta = -2\int \theta P(\theta|\mathcal{D})d\theta + 2\hat{\theta}$
- setting that equal to zero yields $\hat{\theta} = \int \theta P(\theta|\mathcal{D})d\theta = E[\theta|\mathcal{D}]$
- so in other words the optimal action according to squared loss is the posterior mean
- all inferences and actions can be taken with only a prior and loss function
- in the bayesian approach we do not need to justify our estimator, we just need to specify our family of distributions and the prior
- try to use the conjugate prior when you can
- for a lot of data the prior does not matter very much

recap of conditional probabilistic models

conditional probabilistic models

- input space X
- outcome space y
- action space $A = \{P(y)|$p is a probablity distribution on y $\}$
- hypothesis space $\mathcal{F}$ contains prediction function mapping $f : X \rightarrow A$
- prediction function $f \in \mathcal{F} : f(x)$ produces a distribution on y
- a parametric family of conditional densities is a set $\{P(y|x, \theta) : \theta \in \Theta\}$
- where $P(y|x, \theta)$ is a density on the outcome space y for each x in input space and θ is a parameter in the parameter space
- the action space here is a probability distribution not just a decision

likelihood function

- we can as always find our likelihood function given our data set $\boldsymbol{P(\mathcal{D}|x_1...x_n,\theta) = \Pi_{i=1}^{n} P(y_i|x_i,\theta)}$
- the mle estimator is the the $\boldsymbol{\hat{\theta}_{mle} = argmax_{\theta\in\Theta}\mathcal{L}_{\mathcal{D}}(\Theta)}$
- the corresponding prediction function is $\boldsymbol{f(x) = P(y|x,\hat{\theta}_{mle})}$

bayesian condtiional models

- input space $\boldsymbol{X = \mathbb{R}^D}$ outcome space $\boldsymbol{Y = \mathbb{R}}$
- parametric family of distributions $\boldsymbol{\{P(y|x,\theta) : \theta \in \Theta\}}$
- prior $\boldsymbol{P(\theta) : \theta \in \Theta}$ (so we have added a prior)
- the posterior is $\boldsymbol{P(\theta|\mathcal{D},X) \propto P(\mathcal{D}|\theta)P(\theta) = \mathcal{L}_{\mathcal{D}}(\theta)p(\theta)}$
- we don't worry about the denominator in this case because the dataset is fixed so we can just look at the proportional
- we can use bayesian decisions theory to derive point estimates, we have a few choices for how we do this. also depends on our loss function

bayesian prediction function

- we want to find a prediction function that takes input $\boldsymbol{x \in X}$ and produces a distribution on Y
- in the frequentest approach we chose a conditional family of probability distributions (hypothesis space) and select one conditional probability from the family based on some rule like the mle
- in Bayesian setting we chose a parametric family of conditional densities

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

  and a prior distribution $\boldsymbol{P(\theta) : \theta \in \Theta}$
- with a bayesian model how do we predict a distribution on y for input x
- we do not need to make a discrete selection from the hypothesis space: we can maintain some uncertainty

prior predictive distribution

- suppose we have not yes observed any data
- in teh Bayesian setting we can still produce a prediction function
- call **the prior predictive distribution**

$$\boldsymbol{x \rightarrow P(y|x) = \int P(y|x,\theta)p(\theta)d\theta}$$

- this is an average of all conditional densities in our family weighted by the prior.
- so we are considering all possible $\boldsymbol{\theta}$ and we get out a $\boldsymbol{P(y|x)}$

posterior predictive distribution

- after seeing our data set $\boldsymbol{\mathcal{D}}$
- the posterior predictive distribution is given by

$$\boldsymbol{x \rightarrow P(y|x,\mathcal{D}) = \int P(y|x,\theta)P(\theta|\mathcal{D})d\theta}$$

- this is an average of all conditional densities in our hypothesis space weighted by the posterior distribution , so we are again considering all $\boldsymbol{\theta}$
- ewe have not chosen a particular $\boldsymbol{\theta}$ we consider all and just weighted by there likelihood

comparing to the frequentest approach

- in bayesian stats we have two distributions on $\boldsymbol{\Theta}$ the prior, and the posterior $\boldsymbol{P(\theta|\mathcal{D})}$
- there distributions are over parameters corresponding to the distribution on the hypothesis space so, what we get out is a distribution on the hypothesis space
- in the frequentest approach we just pick one $\boldsymbol{\hat{\theta}}$ that we think is best
- in the bayesian approach we weight over all possible outcomes

what if we don't want a full distribution on y

- once we have a predictive distribution $\boldsymbol{P(y|x,\mathcal{D})}$ we can generate a single prediction
- there are many choices depending on what loss we want to minimize

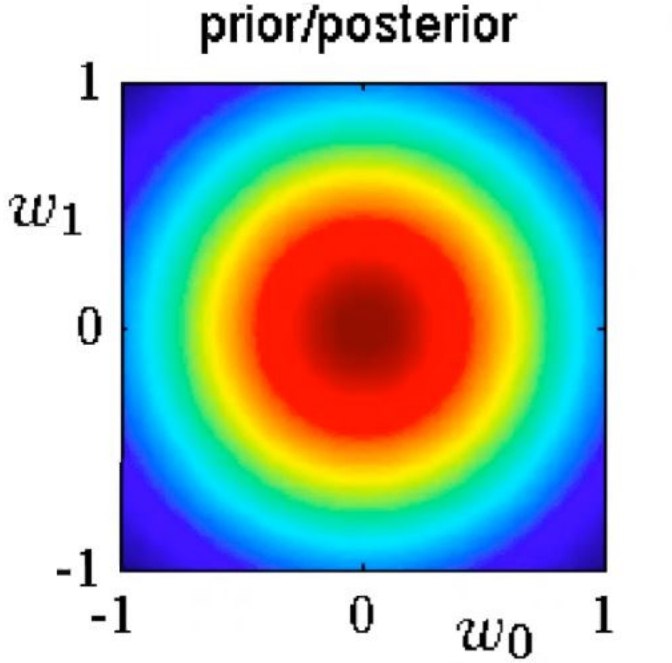
gaussian regression example

example in 1 dimension

- input space $\boldsymbol{X} = [-\mathbf{1}, \mathbf{1}]$ output space $\boldsymbol{y} \in \mathbb{R}$
- y is generated as $\boldsymbol{y} = \boldsymbol{w_0} + \boldsymbol{w_1 x} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \in \boldsymbol{\mathcal{N}(0, .2^2)} \iff \boldsymbol{y|x, w_0, w_1} \sim \boldsymbol{\mathcal{N}(w_0 + w_1 x, .2^2)}$
- lets assume our prior is $\boldsymbol{w} = \boldsymbol{(w_0, w_1)} \sim \boldsymbol{\mathcal{N}(0, \frac{1}{2} * I)}$ where I is the identity matrix
- the parameter space is $\boldsymbol{W} = \mathbb{R}^{\mathbf{2}}$

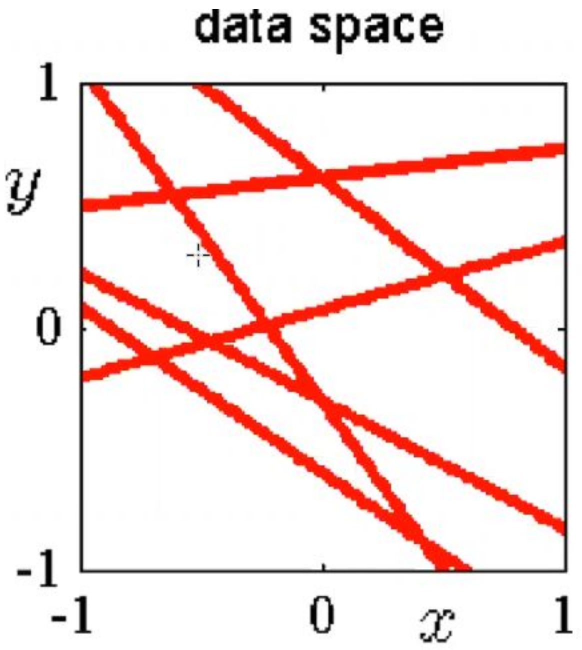
example in 1d prior distribution

- we know that our covariance matrix is $\boldsymbol{\frac{1}{2}I} = \begin{pmatrix} \mathbf{.5} & \mathbf{0} \\ \mathbf{0} & \mathbf{.5} \end{pmatrix}$ so our contour lines are perfect circles radiating out from the mean zero



- so if we were to sample $\boldsymbol{w}$ according to our prior we could basically get any line
- so without any data $\boldsymbol{E[y|x, w]} = \boldsymbol{w_0} + \boldsymbol{w_1 x}$ for a w sampled from $\boldsymbol{w} \sim \boldsymbol{\mathcal{N}(0, \frac{1}{2}I)}$

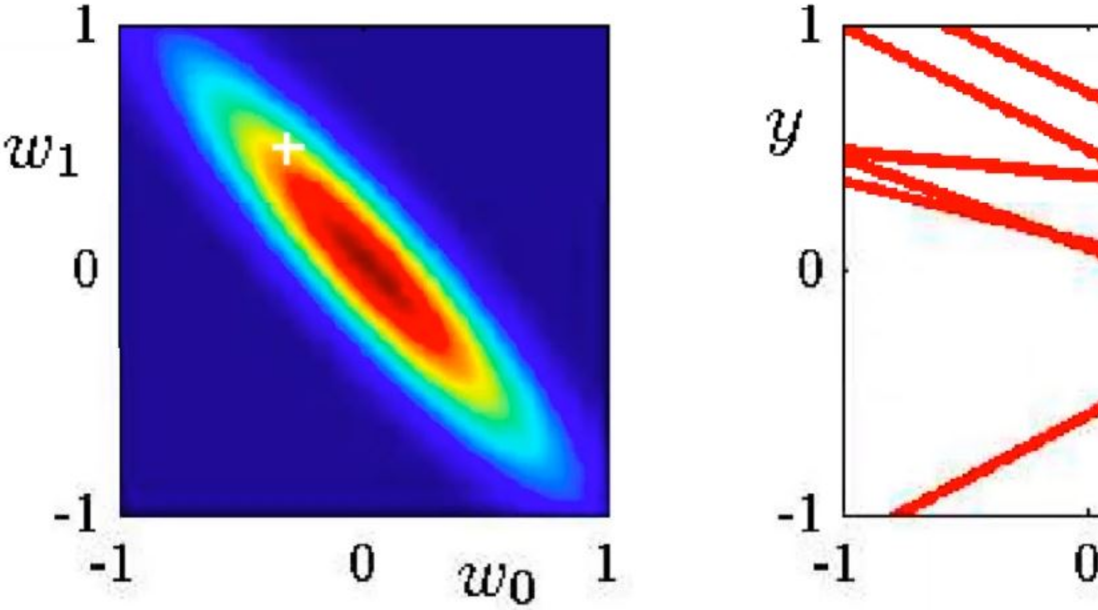
- so the expectation of some random realizations of our prior could look like this



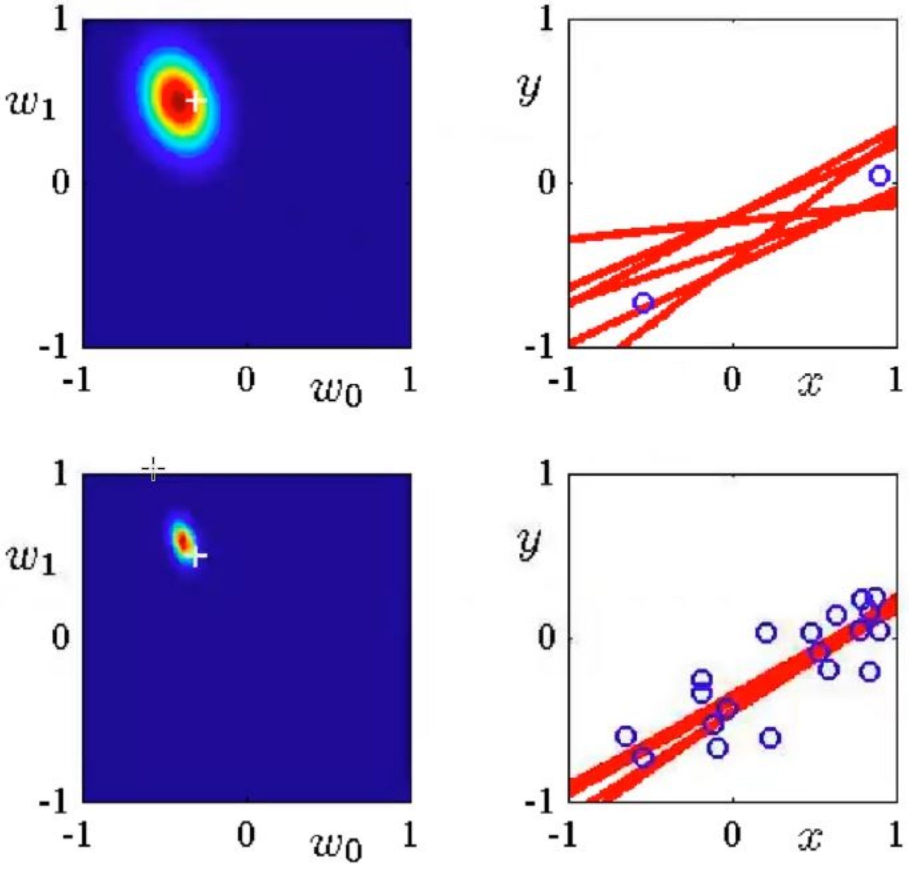
-

example in 1d 1 observation

- if we are then given one observation



-
- lets say it is the blue point on the right graph
- the left is our posterior $P(\theta|\mathcal{D}) = P(\mathcal{D}|\theta)P(\theta)$ so our prior shifts a lot to acuminated the new point
- the lines on the right graph are now $y(x) = E[y|x, w] = w_0 + w_1 x$ but our w is sampled from our posterior $w \sim P(w|\mathcal{D})$
- as we get more data our beliefs collapse into a smaller distribution
-



-

closed from of the posterior

- this is a $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{d}}$ general gaussian linear regression problem
- the gaussian has a conjugate prior so its posterior is also gaussian which is nice
- we model $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma_0})$ and $\boldsymbol{y_i}|\boldsymbol{x}, \boldsymbol{w}$ iid $\mathcal{N}(\boldsymbol{w^t x_i}, \boldsymbol{\sigma^2})$
- we have a design matrix X and response vector y
- the posterior distribution is a guassian distributions with $\boldsymbol{W}|\boldsymbol{\mathcal{D}} \sim \mathcal{N}(\boldsymbol{\mu_p}.\boldsymbol{\Sigma_p})$
- where $\boldsymbol{\mu_p} = (\boldsymbol{X^T X} + \boldsymbol{\sigma^2 \Sigma_0^{-1}})^{-1} \boldsymbol{X^t y}$
- and $\boldsymbol{\Sigma_o} = (\boldsymbol{\sigma^{-}2 X^t X} + \boldsymbol{\Sigma_0^{-1}})^{-1}$

- there parameters tell us where our parameter is centered and how much uncertainty we have about the parameter space

- they skipped how to derive this in class and i am not sure it is super important beyond algebra

- if we want a map point estimate of w then $\boldsymbol{\hat{w} = \mu_p}$

- further we can note that the map estimate with a prior of $\boldsymbol{\Sigma_0 = \frac{\sigma^2}{\lambda} I}$ yields $\boldsymbol{\hat{w} = \mu_p = (X^T X + \lambda I)^{-1} X^T y}$ which is the closed from solution to ridge regression so ridge can be though of as a point estimate gaussian regression with a certain value of $\boldsymbol{\Sigma_0}$

- but the gaussian also lets you understand the overall distribution of your parameters while ridge regression only gives you a single estimate