

ECE-271A
Statistical Learning I:
Bayesian parameter
estimation

Nuno Vasconcelos

ECE Department, UCSD

Bayesian parameter estimation

- ▶ the main difference with respect to ML is that in the Bayesian case Θ is a random variable
- ▶ basic concepts
 - training set $\mathcal{D} = \{x_1, \dots, x_n\}$ of examples drawn independently
 - probability density for observations given parameter

$$P_{X|\Theta}(x | \theta)$$

- prior distribution for parameter configurations

$$P_\Theta(\theta)$$

that encodes prior beliefs about them

- ▶ goal: to compute the posterior distribution

$$P_{\Theta|X}(\theta | D)$$

Bayesian BDR

► pick i if

$$i^*(x) = \arg \max_i P_{X|Y,T}(x | i, D_i) P_Y(i)$$

$$\text{where } P_{X|Y,T}(x | i, D_i) = \int P_{X|Y,\Theta}(x | i, \theta) P_{\Theta|Y,T}(\theta | i, D_i) d\theta$$

► note:

- BDR accounts for ALL information available in the training set
- as before the bottom equation is repeated for each class
- hence, we can drop the dependence on the class
- and consider the more general problem of estimating

$$P_{X|T}(x | D) = \int P_{X|\Theta}(x | \theta) P_{\Theta|T}(\theta | D) d\theta$$

The predictive distribution

► the distribution

$$P_{X|T}(x | D) = \int P_{X|\Theta}(x | \theta) P_{\Theta|T}(\theta | D) d\theta$$

is known as the **predictive distribution**

► note that it can also be written as

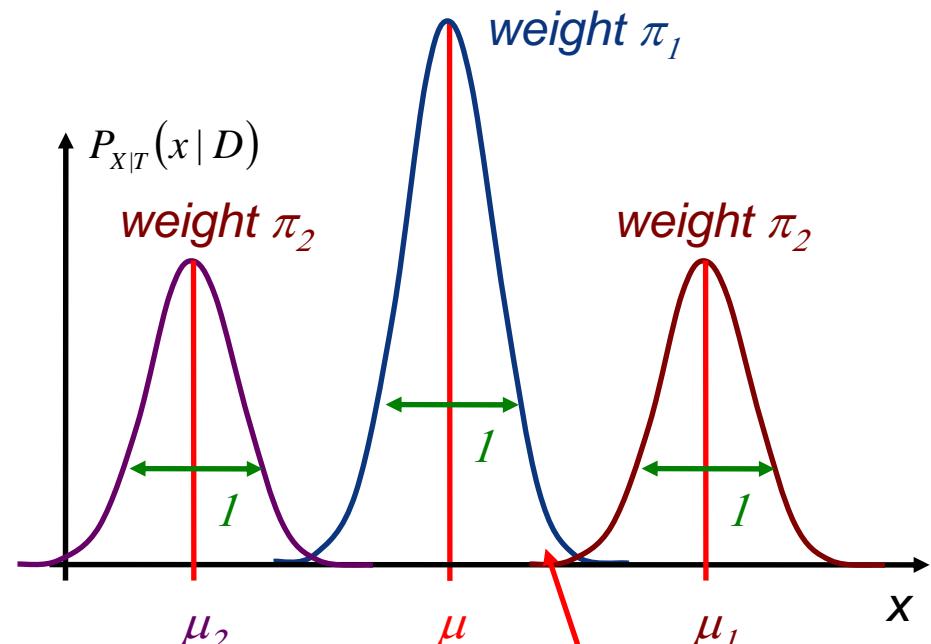
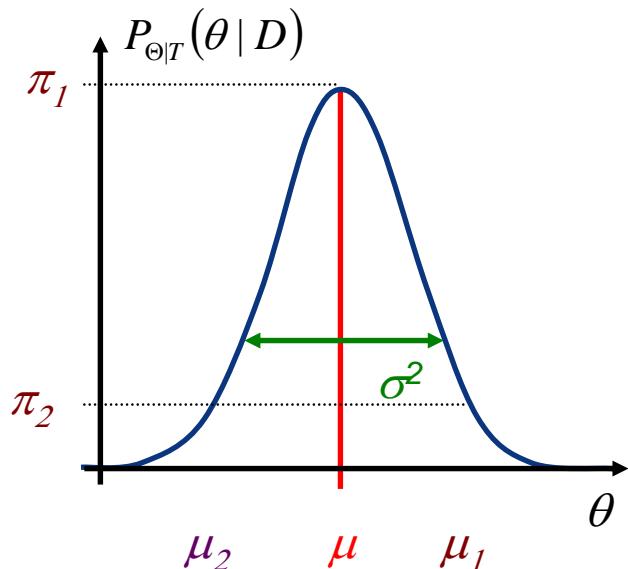
$$P_{X|T}(x | D) = E_{\Theta|T}[P_{X|\Theta}(x | \theta) | T = D]$$

- since **each parameter value defines a model**
- this is an expectation over all possible models
- each **model is weighted by its posterior probability**, given training data

The predictive distribution

► suppose that

$$P_{X|\Theta}(x|\theta) \sim N(\theta, 1) \quad \text{and} \quad P_{\Theta|T}(\theta|D) \sim N(\mu, \sigma^2)$$



► the predictive distribution is an average of all these Gaussians

$$P_{X|T}(x|D) = \int P_{X|\Theta}(x|\theta)P_{\Theta|T}(\theta|D)d\theta$$

MAP vs ML

► ML-BDR

- pick i if

$$i^*(x) = \arg \max_i P_{X|Y}(x | i; \theta_i^*) P_Y(i)$$

$$\text{where } \theta_i^* = \arg \max_{\theta} P_{X|Y}(D | i, \theta)$$

► Bayes MAP-BDR

- pick i if

$$i^*(x) = \arg \max_i P_{X|Y}(x | i; \theta_i^{MAP}) P_Y(i)$$

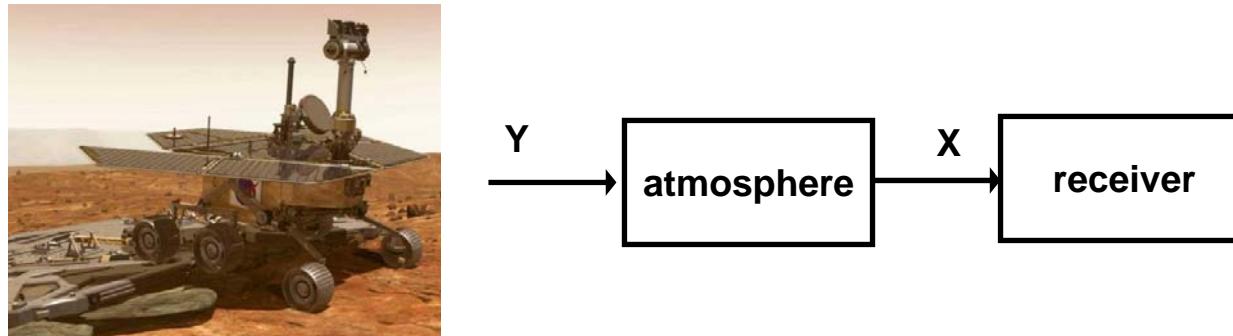
$$\text{where } \theta_i^{MAP} = \arg \max_{\theta} P_{T|Y,\Theta}(D | i, \theta) P_{\Theta|Y}(\theta | i)$$

- the difference is non-negligible only when the dataset is small

► there are better alternative approximations

Example

► communications problem



► two states:

- $Y=0$ transmit signal $s = -\mu_0$
- $Y=1$ transmit signal $s = \mu_0$

► noise model

$$X = Y + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Example

► the BDR is

- pick “0” if

$$x < \frac{\mu_0 + (-\mu_0)}{2} = 0$$

► this is optimal and everything works wonderfully, but

- one day we get a phone call: the receiver is generating a lot of errors!
- there is a calibration mode:
 - rover can send a test sequence
 - but it is expensive, can only send a few bits
- if everything is normal, received means should be μ_0 and $-\mu_0$

Example

► action:

- ask the system to transmit a few 1s and measure X
- compute the ML estimate of the mean of X

$$\mu = \frac{1}{n} \sum_i X_i$$

► result: the estimate is different than μ_0

► we need to combine two forms of information

- our prior is that

$$\mu \sim N(\mu_0, \sigma^2)$$

- our “data driven” estimate is that

$$X \sim N(\hat{\mu}, \sigma^2)$$

Bayesian solution

► Gaussian likelihood (observations)

$$P_{T|\mu}(D | \mu) = G(D, \mu, \sigma^2) \quad \sigma^2 \text{ is known}$$

► Gaussian prior (what we know)

$$P_\mu(\mu) = G(\mu, \mu_0, \sigma_0^2)$$

- μ_0, σ_0^2 are known hyper-parameters

► we need to compute

- posterior distribution for μ

$$P_{\mu|T}(\mu | D) = \frac{P_{T|\mu}(D | \mu)P_\mu(\mu)}{P_T(D)}$$

Bayesian solution

► the posterior distribution is

$$P_{\mu|T}(\mu | D) = G\left(\mu, \mu_n, \sigma_n^2\right)$$

$$\mu_n = \frac{\sigma_0^2 \sum_i x_i + \mu_0 \sigma^2}{\sigma^2 + n\sigma_0^2} \Rightarrow \mu_n = \underbrace{\frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2} \mu_{ML}}_{\alpha_n} + \underbrace{\frac{\sigma^2}{\sigma^2 + n\sigma_0^2} \mu_0}_{1-\alpha_n}$$

$$\sigma_n^2 = \left(\frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2} \right) \Rightarrow \frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

► this is intuitive

Bayesian solution

► for free, Bayes also gives us

- the weighting constants

$$\alpha_n = \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

- a measure of the uncertainty of our estimate

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

- note that $1/\sigma^2$ is a measure of precision
- this should be read as

$$P_{\text{Bayes}} = P_{\text{ML}} + P_{\text{prior}}$$

- Bayesian precision is greater than both that of ML and prior

Observations

- 1) note that precision increases with n, variance goes to zero

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

we are guaranteed that in the limit of infinite data we have convergence to a single estimate

- 2) for large n the likelihood term dominates the prior term

$$\mu_n = \alpha_n \hat{\mu} + (1 - \alpha_n) \mu_0$$
$$\alpha_n \in [0,1], \quad \alpha_n \xrightarrow{n \rightarrow \infty} 1, \quad \alpha_n \xrightarrow{n \rightarrow 0} 0$$

the solution is equivalent to that of ML

- for small n, the prior dominates
- this always happens for Bayesian solutions

$$P_{\mu|T}(\mu | D) \propto \prod_i P_{X|\mu}(x_i | \mu) P_\mu(\mu)$$

Observations

- 3) for a given n

$$\alpha_n = \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

$$\mu_n = \alpha_n \hat{\mu} + (1 - \alpha_n) \mu_0$$
$$\alpha_n \in [0,1], \quad \alpha_n \xrightarrow{n \rightarrow \infty} 1, \quad \alpha_n \xrightarrow{n \rightarrow 0} 0$$

if $\sigma_0^2 \gg \sigma^2$, i.e. we really don't know what μ is a priori
then $\mu_n = \mu_{ML}$

- on the other hand, if $\sigma_0^2 \ll \sigma^2$, i.e. we are very certain a priori,
then $\mu_n = \mu_0$

► in summary,

- Bayesian estimate combines the prior beliefs with the evidence provided by the data
- in a very intuitive manner

Regularization

► regularization:

- if $\sigma_0^2 = \sigma^2$ then $\mu_n = \frac{n}{n+1} \hat{\mu}_{ML} + \frac{1}{n+1} \mu_0$
 $= \frac{1}{n+1} \sum_{i=1}^{n+1} X_i, \quad \text{with } X_{i+1} = \mu_0$

► Bayes is equal to ML on a virtual sample with extra points

- in this case, one additional point equal to the mean of the prior
- for large n, extra point is irrelevant
- for small n, it regularizes the Bayes estimate by
 - directing the posterior mean towards the prior mean
 - reducing the variance of the posterior
$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}$$

► HW: this interpretation holds for all conjugate priors

Conjugate priors

► note that

- the prior $P_\mu(\mu) = G(\mu, \mu_0, \sigma_0^2)$ is Gaussian
- the posterior $P_{\mu|T}(\mu | D) = G(x, \mu_n, \sigma_n^2)$ is Gaussian

► whenever this is the case (posterior in the same family as prior) we say that

- $P_\mu(\mu)$ is a conjugate prior for the likelihood $P_{X|\mu}(x | \mu)$
- posterior $P_{\mu|T}(\mu | D)$ is the reproducing density

► HW: a number of likelihoods have conjugate priors

Likelihood	Conjugate prior
Bernoulli	Beta
Poisson	Gamma
Exponential	Gamma
Normal (known σ^2)	Gamma

Exponential family

- ▶ you will also show that all of these likelihoods are members of the exponential family

$$P_{X|\Theta}(x | \theta) = f(x)g(\theta) e^{\phi(\theta)^T u(x)}$$

- ▶ for this family, the interpretation of Bayesian parameter estimation as “ML on a properly augmented sample” always holds (whenever the prior is the conjugate)
- ▶ this is one of the reasons why the exponential family is “special” (but there are others)

Predictive distribution

- ▶ we have seen that $P_{\mu|T}(\mu | D) = G(x, \mu_n, \sigma_n^2)$
- ▶ we can now compute the predictive distribution

$$\begin{aligned} P_{X|T}(x | D) &= \int P_{X|\mu}(x | \mu) P_{\mu|T}(\mu | D) d\mu \\ &= \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(\mu-\mu_n)^2}{2\sigma_n^2}} d\mu \end{aligned}$$

$$= \int f(x - \mu) h(\mu) d\mu$$

$$\left(\text{with } f(x) = G(x, 0, \sigma^2) \text{ and } h(x) = G(x, \mu_n, \sigma_n^2) \right)$$
$$= G(x, 0, \sigma^2) * G(x, \mu_n, \sigma_n^2)$$

- ▶ i.e. $X|T$ is the random variable that results from adding two independent Gaussians with these parameters

Predictive distribution

► hence $X|T$ is Gaussian with

$$P_{X|T}(x | D) = G(x, \mu_n, \sigma^2 + \sigma_n^2)$$

- the mean is that of the posterior
- variance increased by σ^2 to account for the uncertainty of the observations

► note:

- we will not go over the multivariate case in class, but the expressions are straightforward generalization
- make sure you are comfortable with them

Priors

- ▶ potential problem of the Bayesian framework
 - “I don’t really have a strong belief about what the most likely parameter configuration is”
- ▶ in these cases it is usual to adopt a non-informative prior
- ▶ the most obvious choice is the uniform distribution

$$P_{\Theta}(\theta) = \alpha$$

- ▶ there are, however, problems with this choice

- if θ is unbounded this is an improper distribution

$$\int_{-\infty}^{\infty} P_{\Theta}(\theta) d\theta = \infty \neq 1$$

- the prior is not invariant to all reparametrizations

Example

- ▶ consider Θ and a new random variable η with $\eta = e^\theta$
- ▶ since this is a 1-to-1 transformation it should not affect the outcome of the inference process
- ▶ we check this by using the change of variable theorem
 - if $y = f(x)$ then

$$P_Y(y) = \frac{1}{\left| \frac{\partial f}{\partial X} \right|_{x=f^{-1}(y)}} P_X(f^{-1}(y))$$

- ▶ in this case

$$P_\eta(\eta) = \frac{1}{\left| \frac{\partial e^\theta}{\partial \theta} \right|_{\theta=\log \eta}} P_\Theta(\log \eta) = \frac{1}{|\eta|} P_\Theta(\log \eta)$$

Invariant non-informative priors

- ▶ for uniform θ this means that $P_\eta(\eta) \propto \frac{1}{|\eta|}$, i.e. not constant
- ▶ this means that
 - there is no consistency between θ and h
 - a 1-to-1 transformation changes the non-informative prior into an informative one
- ▶ to avoid this problem the non-informative prior has to be invariant
- ▶ e.g. consider a location parameter:
 - a parameter that simply shifts the density
 - e.g. the mean of a Gaussian
- ▶ a non-informative prior for a location parameter has to be invariant to shifts, i.e. the transformation $Y = \mu + c$

Location parameters

► in this case

$$P_Y(y) = \frac{1}{\left| \frac{\partial(\mu + c)}{\partial \mu} \right|_{\mu=y-c}} P_\mu(y - c) = P_\mu(y - c)$$

and, since this has to be valid for all c ,

$$P_Y(y) = P_\mu(y)$$

► hence

$$P_\mu(y - c) = P_\mu(y)$$

- which is valid for all c if and only if $P_\mu(\mu)$ is uniform
- non-informative prior for location is $P_\mu(\mu) \propto 1$

Scale parameters

- ▶ a scale parameter is one that controls the scale of the density

$$\sigma^{-1} f\left(\frac{x}{\sigma}\right)$$

e.g. the variance of a Gaussian distribution

- ▶ it can be shown that, in this case, the non-informative prior invariant to scale transformations is

$$P_\sigma(\sigma) = \frac{1}{\sigma}$$

- ▶ note that, as for location, this is an improper prior

Selecting priors

- ▶ non-informative priors are the end of the spectrum where we don't know what parameter values to favor
- ▶ at the other end, i.e. when we are absolutely sure, the prior becomes a delta function

$$P_{\Theta}(\theta) = \delta(\theta - \theta_0)$$

- ▶ in this case

$$P_{\Theta|T}(\theta | D) \propto P_{T|\Theta}(D | \theta) \delta(\theta - \theta_0)$$

and the predictive distribution is

$$\begin{aligned} P_{X|T}(x | D) &\propto \int P_{X|\Theta}(x | \theta) P_{T|\Theta}(D | \theta) \delta(\theta - \theta_0) d\theta \\ &= P_{X|\Theta}(x | \theta_0) \end{aligned}$$

- ▶ this is identical to ML if $\theta_0 = \theta_{ML}$

Selecting priors

- ▶ hence,
 - ML is a special case of the Bayesian formulation,
 - where we are absolutely confident that the ML estimate is the correct value for the parameter
 - ▶ but we could use other values for θ_0 . For example the value that maximizes the posterior
- $$\theta_{MAP} = \arg \max_{\theta} P_{\Theta|T}(\theta | D) = \arg \max_{\theta} P_{T|\Theta}(D | \theta)P_{\Theta}(\theta)$$
- ▶ this is called the MAP estimate and makes the predictive distribution equal to
- $$P_{X|T}(x | D) = P_{X|\Theta}(x | \theta_{MAP})$$
- ▶ it can be useful when the true predictive distribution has no closed-form solution

Selecting priors

► the natural question is then

- “what if I don’t get the prior right?”; “can I do terribly bad?”
- “how robust is the Bayesian solution to the choice of prior?”
- let’s see how much the solution changes between the two extremes

► for the Gaussian problem

- absolute certainty priors: $P_\mu(\mu) = \delta(\mu - \mu_p)$
 - MAP estimate: since $P_{\mu|T}(\mu | D) = G(x, \mu_n, \sigma_n^2)$ we have

$$\mu_p = \mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \mu_{ML} + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0$$

- ML estimate is $\mu_p = \mu_{ML}$
- we have seen already that these are similar unless the sample is small (MAP = ML on sample with extra point)

Selecting priors

► for the Gaussian problem

- non-informative prior:
 - in this case it is $P_\mu(\mu) \propto 1$ or

$$P_\mu(\mu) = \lim_{\sigma_0^2 \rightarrow \infty} G(\mu, \mu_0, \sigma_0^2)$$

- from which

$$\mu_n = \lim_{\sigma_0^2 \rightarrow \infty} \left(\frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \mu_{ML} + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 \right) = \mu_{ML}$$

$$\frac{1}{\sigma_n^2} = \lim_{\sigma_0^2 \rightarrow \infty} \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) = \frac{n}{\sigma^2} \Leftrightarrow \sigma_n^2 = \sigma_{ML}^2$$

- and

$$P_{X|T}(x | D) = G(x, \mu_n, \sigma^2 + \sigma_n^2) = G\left(x, \mu_{ML}, \sigma^2 \left(1 + \frac{1}{n}\right)\right)$$

Selecting priors

► in summary, for the two prior extremes

- delta prior centered on MAP:

$$P_{x|T}(x | D) = G(x, \mu_{MAP}, \sigma^2)$$

$$\mu_{MAP} = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \mu_{ML} + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0$$

- delta prior centered on ML:

$$P_{x|T}(x | D) = G(x, \mu_{ML}, \sigma^2)$$

- non-informative prior

$$P_{x|T}(x | D) = G\left(x, \mu_{ML}, \sigma^2 \left(1 + \frac{1}{n}\right)\right)$$

► all Gaussian, “qualitatively the same”:

- somewhat different parameters for small n; equal for large n

► this indicates robustness to “incorrect” priors!

Any questions?