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 $\mathcal{D} = \{x_1, ..., x_n\}$ of examples drawn independently
  - probability density for observations given parameter

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

  - prior distribution for parameter configurations

$$\boxed{P_{\Theta}(\theta)}$$

  that encodes prior beliefs about them

- goal: to compute the posterior distribution

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

Bayesian BDR

- pick i if

$$\boxed{\begin{aligned} &i^*(x) = \arg\max_i P_{X|Y,T}\left(x \mid i, D_i\right) P_Y(i) \\ &where \quad P_{X|Y,T}\left(x \mid i, D_i\right) = \int P_{X|Y,\Theta}\left(x \mid i, \theta\right) P_{\Theta|Y,T}\left(\theta \mid i, D_i\right) d\theta \end{aligned}}$$

- 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

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

The predictive distribution

- the distribution

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

  is known as the predictive distribution

- note that it can also be written as

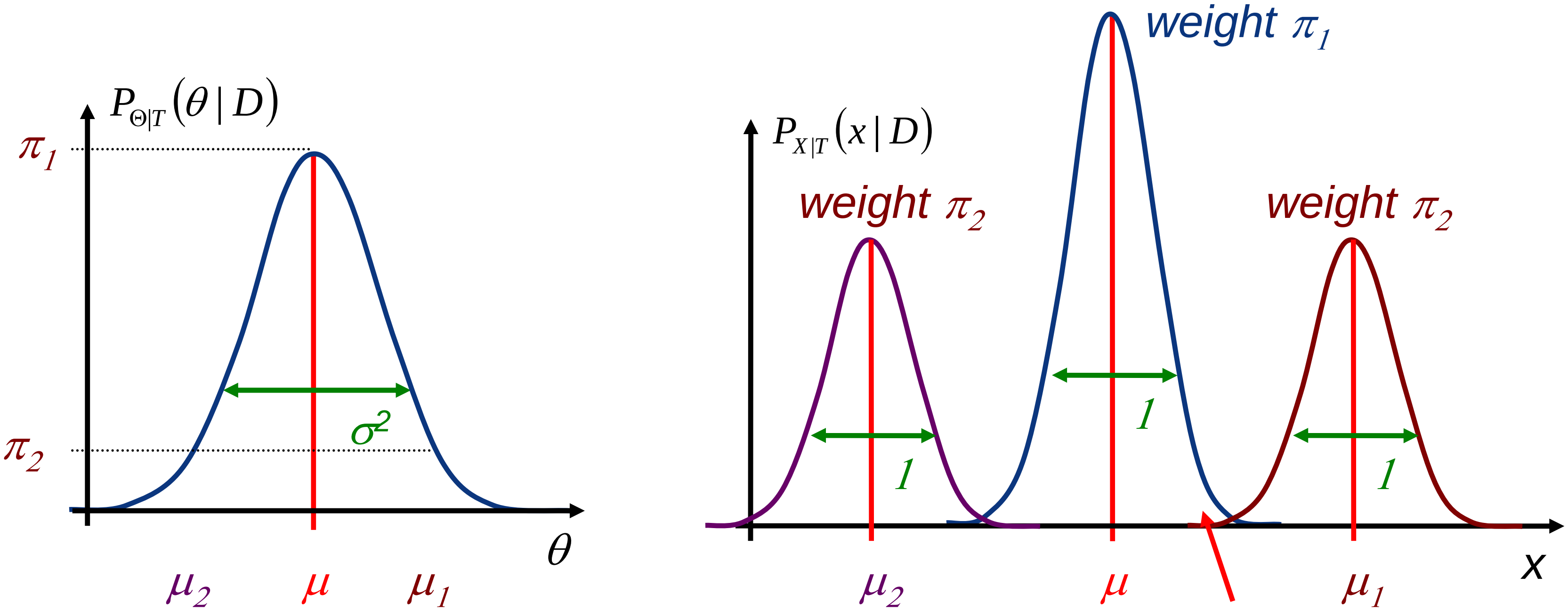
$$\boxed{P_{X|T}(x \mid D) = E_{\Theta|T}\left[P_{X|\Theta}(x \mid \theta) \mid T = D\right]}$$

  - 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

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



- the predictive distribution is an average of all these Gaussians

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

MAP vs ML

- ML-BDR
  - pick i if

$$\boxed{\begin{aligned} &i^*(x) = \arg\max_i P_{X|Y}\left(x \mid i;\theta_i^*\right) P_Y(i) \\ &\text{where } \theta_i^* = \arg\max_\theta P_{X|Y}\left(D \mid i,\theta\right) \end{aligned}}$$

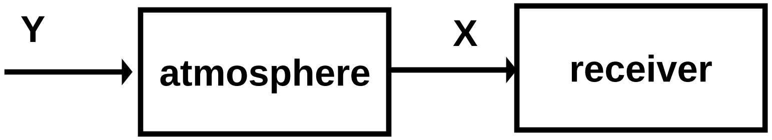
- Bayes MAP-BDR
  - pick i if

$$\boxed{\begin{aligned} &i^*(x) = \arg\max_i P_{X|Y}\left(x \mid i;\theta_i^{MAP}\right) P_Y(i) \\ &\text{where } \theta_i^{MAP} = \arg\max_\theta P_{T|Y,\Theta}\left(D \mid i,\theta\right) P_{\Theta|Y}\left(\theta \mid i\right) \end{aligned}}$$

  - 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 = μ_0
- noise model

$$\boxed{X = Y + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)}$$

Example

- the BDR is
  - pick "0" if

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

- 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

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

- result: the estimate is different than μ_0
- we need to combine two forms of information
  - our prior is that $\boxed{\mu \sim N(\mu_0, \sigma^2)}$
  - our "data driven" estimate is that

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

Bayesian solution

- Gaussian likelihood (observations)

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

- Gaussian prior (what we know)

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

  - μ_0, σ_0^2 are known hyper-parameters

- we need to compute
  - posterior distribution for μ

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

Bayesian solution

- the posterior distribution is

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

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

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

- this is intuitive

Bayesian solution

- for free, Bayes also gives us
  - the weighting constants

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

  - a measure of the uncertainty of our estimate

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

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

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

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

Observations

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

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

  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

$$\boxed{\begin{array}{l} \mu_n = \alpha_n \hat{\mu} + (1 - \alpha_n)\mu_0 \\ \qquad \alpha_n \in [0,1], \quad \alpha_n \underset{n \to \infty}{\longrightarrow} 1, \quad \alpha_n \underset{n \to 0}{\longrightarrow} 0 \end{array}}$$

  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

$$\boxed{\alpha_n = \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}} \quad \boxed{\begin{array}{l} \mu_n = \alpha_n \hat{\mu} + (1-\alpha_n)\mu_0 \\ \qquad \alpha_n \in [0,1], \quad \alpha_n \underset{n\to\infty}{\longrightarrow} 1, \quad \alpha_n \underset{n\to 0}{\longrightarrow} 0 \end{array}}$$

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

- on the other hand, if $\sigma_0^2 << \sigma^2$, 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 $P_\mu(\mu) = G\left(\mu, \mu_0, \sigma_0^2\right)$ is Gaussian
  - the posterior $P_{\mu|T}(\mu \mid D) = G\left(x, \mu_n, \sigma_n^2\right)$ 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 $P_{X|\mu}(x \mid \mu)$
  - posterior $P_{\mu|T}(\mu \mid D)$ 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

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

- 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 $P_{\mu|T}(\mu \mid D) = G\left(x, \mu_n, \sigma_n^2\right)$
- we can now compute the predictive distribution

$$P_{X|T}(x \mid D) = \int P_{X|\mu}(x \mid \mu) P_{\mu|T}(\mu \mid D) d\mu$$

$$= \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(\mu-\mu_n)^2}{2\sigma_n^2}} d\mu$$

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

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

$$= G(x, 0, \sigma^2) * G(x, \mu_n, \sigma_n^2)$$

- 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

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

  - 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

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

- 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

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

- in this case

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

Invariant non-informative priors

- for uniform θ this means that $P_\eta(\eta) \alpha \frac{1}{|\eta|}$, 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

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

- 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

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

- in this case

$$P_{\Theta|T}(\theta \mid D) \;\alpha\; P_{T|\Theta}(D \mid \theta)\delta(\theta - \theta_0)$$

and the predictive distribution is

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

- 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

$$\boxed{\theta_{MAP} = \arg\max_{\theta} P_{\Theta|T}(\theta \mid D) = \arg\max_{\theta} P_{T|\Theta}(D \mid \theta) P_{\Theta}(\theta)}$$

- this is called the MAP estimate and makes the predictive distribution equal to

$$\boxed{P_{X|T}(x \mid D) = P_{X|\Theta}(x \mid \theta_{MAP})}$$

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

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

    - 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 $P_\mu(\mu) \; \alpha \; 1$ or

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

    - from which

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

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

    - and

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

Selecting priors

- in summary, for the two prior extremes
  - delta prior centered on MAP:

$$\boxed{P_{X|T}(x \mid D) = G\left(x, \mu_{MAP}, \sigma^2\right)}$$

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

  - delta prior centered on ML:

$$\boxed{P_{X|T}(x \mid D) = G\left(x, \mu_{ML}, \sigma^2\right)}$$

  - non-informative prior

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

- all Gaussian, "qualitatively the same":
  - somewhat different parameters for small n; equal for large n
- this indicates robustness to "incorrect" priors!

Any questions?