ECE-271A
Statistical Learning I:
Bayesian parameter estimation

Nuno Vasconcelos

ECE Department, UCSD

Bayesian estimation

- last class we considered the Gaussian problem

$$P_{X|\mu}(x|\mu) = G\left(x,\mu,\sigma^2\right), \ \sigma^2 \text{ known} \qquad P_{\mu}(\mu) = G\left(\mu,\mu_0,\sigma_0^2\right)$$

and showed that

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

with

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\hat{\mu}_{ML} + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0$$

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

- good example of various properties that are typical of Bayesian parameter estimates

Properties

- regularization:
  - if $\sigma_0^2 = \sigma^2$ then $\mu_n = \frac{n}{n+1}\hat{\mu}_{ML} + \frac{1}{n+1}\mu_0$

$$= \frac{1}{n+1}\sum_{i=1}^{n+1} X_i, \qquad \text{with } X_{i+1} = \mu_0$$

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

Conjugate priors

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

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 variables 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 η 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

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

- 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

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

and the predictive distribution is

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

- 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

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

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

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

- 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

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

    - 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

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

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

    - and

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

Selecting priors

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

$$P_{X|T}(x|D) = G\left(x, \mu_{MAP}, \sigma^2\right) \qquad \mu_{MAP} = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\mu_{ML} + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

  - delta prior centered on ML:

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

  - 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!

Selecting priors

- another example, problem 3.5.17 DHS (HW prob 3)
  - multivariate Bernoulli (*d* independent Bernoulli variables)
  - since Bernoulli is

$$P_{X|\Theta}(x \mid \theta) = \begin{cases} \theta, & x = 1 \\ 1-\theta, & x = 0 \end{cases} = \theta^{x}(1-\theta)^{1-x}$$

  - multivariate likelihood is:

$$P_{X|\Theta}(x \mid \theta) = \prod_{i=1}^{d} \theta_i^{x_i}(1-\theta_i)^{1-x_i}$$

  - in (a) you show that if $D = \{x^{(1)}, \ldots, x^{(n)}\}$ is a set of *n* iid samples, then

$$P_{T|\Theta}(D \mid \theta) = \prod_{i=1}^{d} \theta_i^{s_i}(1-\theta_i)^{n-s_i}, \qquad s_i = \sum_{j=1}^{n} x_i^{(j)}$$

Selecting priors

- another example, problem 3.5.17 DHS (HW prob 3)
  - in (b) you then show that if Θ is uniform (non-informative) the predictive distribution is

$$P_{X|T}(x \mid D) = \prod_{i=1}^{d} \left(\frac{s_i + 1}{n+2} \right)^{x_i} \left(1 - \frac{s_i + 1}{n+2} \right)^{1-x_i}$$

  - in (d) you show that comparing with

$$P_{X|\Theta}(x \mid \theta) = \prod_{i=1}^{d} \theta_i^{x_i} (1-\theta_i)^{1-x_i}$$

  - this can be interpreted as:
    - under Bayes, with a uniform prior, the predicted distribution is the same as the likelihood, with the parameter estimate

$$\hat{\theta}_i = \frac{s_i + 1}{n+2}$$

Selecting priors

- let's now consider the extreme of $P_\Theta(\theta) = \delta\left(\theta - \hat{\theta}\right)$
  - ML: we know that

$$\hat{\theta}_i = \frac{s_i}{n}$$

  - and

$$P_{X|T}(x \mid D) = \prod_{i=1}^{d} \left(\frac{s_i}{n}\right)^{x_i} \left(1 - \frac{s_i}{n}\right)^{1-x_i}$$

  - this can be interpreted as:
    - the predicted distribution is the same as the likelihood, with the parameter estimate

$$\hat{\theta}_i = \frac{s_i}{n}$$

Selecting priors

- MAP: given prior $P_{\Theta} = \prod_i P_{\Theta_i}(\theta_i)$

$$\hat{\theta} = \arg\max_{\theta} \left\{ \log P_{T|\Theta}(D|\theta) + \log P_{\Theta}(\theta) \right\}$$

- and since

$$P_{T|\Theta}(D|\theta) = \prod_{i=1}^{d} \theta_i^{s_i} \left(1-\theta_i\right)^{n-s_i}, \qquad s_i = \sum_{j=1}^{n} x_i^{(j)}$$

- this is

$$\hat{\theta}_i = \arg\max_{\theta} \left\{ s_i \log\theta_i + (n-s_i)\log(1-\theta_i) + \log P_{\Theta_i}(\theta_i) \right\}$$

- i.e. the solution of

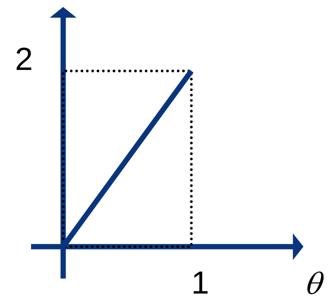
$$\frac{s_i}{\theta_i} - \frac{(n-s_i)}{1-\theta_i} + \frac{1}{P_{\Theta_i}(\theta_i)} \frac{\partial}{\partial\theta_i} P_{\Theta_i}(\theta_i) = 0$$

- let's consider some specific priors

Selecting priors

- prior that favors "1"s

$$P_{\Theta_i}(\theta) = 2\theta$$



- MAP solution:

$$\frac{s_i}{\theta_i} - \frac{(n-s_i)}{1-\theta_i} + \frac{1}{\theta_i} = 0 \quad \Leftrightarrow \quad \hat{\theta}_i = \frac{s_i+1}{n+1}$$

- and

$$P_{X|T}(x|D) = \prod_{i=1}^{d} \left(\frac{s_i+1}{n+1}\right)^{x_i} \left(1-\frac{s_i+1}{n+1}\right)^{1-x_i}$$

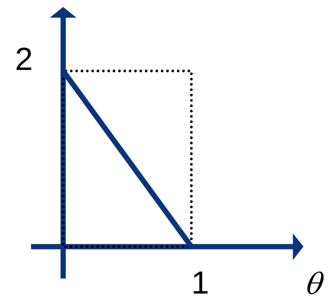
- this can be interpreted as:
  - the predicted distribution is the same as the likelihood, with the parameter estimate

$$\hat{\theta}_i = \frac{s_i+1}{n+1}$$

Selecting priors

- prior that favors "0"s

$$P_{\Theta_i}(\theta) = 2(1-\theta)$$



- MAP solution:

$$\frac{s_i}{\theta_i} - \frac{(n-s_i)}{1-\theta_i} - \frac{1}{1-\theta_i} = 0 \quad \Leftrightarrow \quad \hat{\theta}_i = \frac{s_i}{n+1}$$

- and

$$P_{X|T}(x \mid D) = \prod_{i=1}^{d} \left(\frac{s_i}{n+1}\right)^{x_i} \left(1 - \frac{s_i}{n+1}\right)^{1-x_i}$$

- this can be interpreted as:
  - the predicted distribution is the same as the likelihood, with the parameter estimate

$$\hat{\theta}_i = \frac{s_i}{n+1}$$

Selecting priors

- in summary
  - all cases are of the form $P_{X|T}(x \mid D) = \prod_{i=1}^{d} \hat{\theta}^{x_i} \left(1 - \hat{\theta}\right)^{1-x_i}$
  - with

Estimator	$\hat{\theta}_i$	# tosses	# "1"s	interpretation
ML	s_i / n	n	s_i	
MAP non-informative	s_i / n	n	s_i	"the same"
MAP favor "1"s	$(s_i + 1)/(n + 1)$	$n+1$	s_i+1	"add one 1"
MAP favor "0"s	$s_i /(n + 1)$	$n+1$	s_i	"add one 0"
Bayes non-informative	$(s_i + 1)/(n + 2)$	$n+2$	s_i+1	"add one of each"

  - all cases qualitatively the same: "ML estimate on an extended sample with extra points that reflect the bias of the prior".

Regularization

- these are all examples of regularization
- Q: what is the point of "adding one of each?" by Bayes non-informative?
  - the main problem of ML (s_i / n) is the "empty bin" problem
  - for small n, s_i is likely to be zero independently of the value of θ_i
  - this can lead to all sorts of problems, e.g. a likelihood ratio that goes to infinity
  - by adding "one of each" Bayes eliminates this problem
  - for richly populated bins it makes no difference, but it matters for empty bins
- note that this is consistent with the non-informative prior
  - empty bins are as likely as any other value
  - if we see a lot of them, we need to correct this

Regularization

- “empty bin” problem
  - “why should I care?” this is unlikely if I have a large sample
  - remember that “large” is always relative
  - 10 bins in 1D transforms into 100 in 2D, 1000 in 3D, and 10^d in a d-dimensional space
  - when d is large, we are always in the “small sample” regime
  - regularization usually makes a tremendous difference

- example:
  - histogram estimates in high-dimensional spaces
  - e.g. histogram of English words for indexing web-pages
    - for each page, compute histogram $C = (c_1, \ldots, c_w)$ where c_i is the # of times word i^{th} word appeared in page
    - measure similarity between pages *i,j* with some function $d(C^i,C^j)$

Regularization

- histogram similarity:
  - natural measure is the Kullback-Leibler divergence

$$d(C^i, C^j) = \sum_{k=1}^{W} p_k^i \log\left(\frac{p_k^i}{p_k^j}\right)$$

  - where the probabilities are the counts after normalization

$$p_k^i = c_k^i \Big/ \sum_k c_k^i$$

  - problem: log goes to infinity when $p^j_k = 0$!
  - for low-frequency words the noisy estimates are amplified by the ratio of probabilities
  - the distance measure has a large variance

Regularization

- Prob 3 on HW
  - the count vector C is distributed according to a multinomial distribution

$$P_C(c_1, \ldots, c_W) = \frac{n!}{\prod_{k=1}^{W} c_k!} \prod_{j=1}^{W} \pi_j^{c_j}$$

  - where π_j is the probability of word j.
  - since the π_j are probabilities, we can't use any prior here.
  - distribution over vectors $\pi = (\pi_1, \ldots, \pi_W)$ must satisfy the constraints of a probability mass function

$$\pi_j > 0$$

$$\sum_j \pi_j = 1$$

Regularization

- Prob 3 on HW
  - one such distribution is the Dirichlet distribution

$$P_{\Pi}(\pi_1, \ldots, \pi_W) = \frac{\Gamma\left(\sum_{j=1}^{W} u_j\right)}{\prod_{k=1}^{W} \Gamma(u_j)} \prod_{j=1}^{W} \pi_j^{u_j - 1}$$

  - u_j are hyper-parameters
  - $\Gamma(.)$ is the gamma function

Regularization

- Prob 3 on HW
  - on HW you will show that the posterior is

$$P_{\Pi|C}(\pi | c) = \frac{\Gamma\left(\sum_{j=1}^{W} c_j + u_j\right)}{\prod_{k=1}^{W} \Gamma\left(c_j + u_j\right)} \prod_{j=1}^{W} \pi_j^{c_j + u_j - 1}$$

  - i.e. Dirichlet of hyper-parameters $c_j + u_j$
  - the prior parameters can be seen as additional counts that regularize the predictive distribution!

Any questions?