ECE-271A
Statistical Learning I: Bayesian parameter estimation

Nuno Vasconcelos

ECE Department, UCSD
Bayesian parameter estimation

➤ basic concepts:

• explicitly define the set of training random variables, $T = \{X_1,...,X_n\}$, from which training set $D = \{x_1, ..., x_n\}$ is drawn

• parameter $\Theta$ is a random variable with prior density $P_\Theta(\theta)$ which encodes “observer’s beliefs” on experimental outcomes

• likelihood function $P_{X|\Theta}(X | \theta)$ gives distribution of observations given parameter value

➤ goal:

• estimate the complete posterior distribution $P_{\Theta|T}(\theta | D)$ to obtain a complete characterization of $\Theta$
Bayesian predictions

e.g. class-conditional likelihoods that we use in the BDR

are based on the predictive distribution

\[
P_{X|T}(x \mid D) = \int P_{X|\theta}(x \mid \theta)P_{\theta|T}(\theta \mid D)d\theta
\]

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

- average of all models (unlike ML estimate)
- each model weighted by how much it is plausible, given the training set \(D\)

all predictions conditioned on \(T\)

- no information is lost
- posterior is a compact representation of the training set
- computationally efficient when the integral has closed form
The Gaussian case

last class we considered the Gaussian problem

\[ P_{X|\mu}(x | \mu) = G(x, \mu, \sigma^2), \quad \sigma^2 \text{ known} \]

and showed that

\[ P_{\mu|D}(\mu | D) = G(x, \mu_n, \sigma_n^2) \quad P_{X|D}(x | D) = G(x, \mu_n, \sigma^2 + \sigma_n^2) \]

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

- **likelihood dominates** as \( n \) goes to infinity
- **prior dominates** as \( n \) goes to zero

\[
\mu_n = \alpha_n \hat{\mu}_{ML} + (1 - \alpha_n) \mu_0 \quad \text{with} \quad \alpha_n \to 0 \quad \alpha_n \to 1
\]

- **\( \mu_n \) linear interpolant** between the two solutions
- **variance of posterior** goes to zero as \( n \) goes to infinity

\[
\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \quad \Leftrightarrow \quad \text{prec}_{\text{Bayes}} = \text{prec}_{\text{ML}} + \text{prec}_{\text{prior}}
\]

- **Bayes strictly more precise** than either ML or prior
- **intuitive balance** between prior and likelihood parameters

\[
\sigma_0^2 \ll \sigma^2 \quad \Rightarrow \quad \alpha_n \approx 0 \quad \text{prior dominates}
\]

\[
\sigma_0^2 \gg \sigma^2 \quad \Rightarrow \quad \alpha_n \approx 1 \quad \text{likelihood dominates}
\]
Properties

- **regularization:**
  - if \( \sigma_0^2 = \sigma^2 \) then
    \[
    \mu_n = \frac{1}{n+1} \hat{\mu}_{ML} + \frac{1}{n+1} \mu_0
    \]
    \[
    = \frac{1}{n+1} \sum_{i=1}^{n+1} X_i, \quad \text{with } X_{i+1} = \mu_0
    \]

- 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 $P_{\mu|D}(\mu | D) = G(x, \mu_n, \sigma_n^2)$ 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 $P_{\mu|D}(\mu | D)$ is the reproducing density

HW: a number of likelihoods have conjugate priors

<table>
<thead>
<tr>
<th>Likelihood</th>
<th>Conjugate prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernoulli</td>
<td>Beta</td>
</tr>
<tr>
<td>Poisson</td>
<td>Gamma</td>
</tr>
<tr>
<td>Exponential</td>
<td>Gamma</td>
</tr>
<tr>
<td>Normal (known $\sigma^2$)</td>
<td>Gamma</td>
</tr>
</tbody>
</table>
Exponential family

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

\[ P_X(x | \theta) = f(x)g(\theta) \, e^{\phi(\theta)^T u(x)} \]

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)
Priors

typical 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

- non-informative priors invariant to reparametrization
  
  - for location
    
    \[ P_\mu(\mu) = \mu \]
  
  - for scale
    
    \[ P_\sigma(\sigma) = \frac{1}{\sigma} \]

- note that these are improper distributions
  
  \[ \int_{-\infty}^{\infty} P_\theta(\theta)d\theta = \infty \neq 1 \]
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
  \[
P_{X|T}(x | D) \propto \int P_{X|\Theta}(x | \theta)P_{T|\Theta}(D | \theta)\delta(\theta - \theta_0)d\theta
  = P_{X|\Theta}(x | \theta_0)
  \]
  - 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 $\theta_0$. 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

$$P_{X\mid T}(X \mid D) = P_{X\mid \theta}(X \mid \theta_{MAP})$$

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: \( P_\mu(\mu) = \delta(\mu - \mu_0) \)

- MAP estimate: since \( P_\mu(\mu | D) = G(\mu, \mu_n, \sigma_n^2) \) we have
  \[
  \mu_0 = \mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \mu_{ML} + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0
  \]

- ML estimate is \( \mu_0 = \mu_{ML} \)
Selecting priors

for the Gaussian problem

• non-informative prior:
  • in this case it is $P_{\mu}(\mu) \propto 1$ or
  \[ P_{\mu}(\mu) = \lim_{\sigma_0^2 \to \infty} G(\mu, \mu_0, \sigma_0^2) \]
  • 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:

\[
P_{X|T}(X | D) = G(X, \mu_{\text{MAP}}, \sigma^2)
\]

- Delta prior centered on ML:

\[
P_{X|T}(X | D) = G(X, \mu_{\text{ML}}, \sigma^2)
\]

- Non-informative prior

\[
P_{X|T}(X | D) = G(X, \mu_{\text{ML}}, \sigma^2 \left(1 + \frac{1}{n}\right))
\]

- 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
    \[
    P_{X|\Theta}(x | \theta) = \begin{cases} 
    \theta, & x = 1 \\
    1 - \theta, & x = 0 
    \end{cases} = \theta^x (1 - \theta)^{1-x}
    \]
  - multivariate likelihood is:
    \[
    P_{X|\Theta}(x | \theta) = \prod_{i=1}^{d} \theta_i^{x_i} (1 - \theta_i)^{1-x_i}
    \]
  - in (a) you show that if \(D = \{x^{(1)}, \ldots, x^{(n)}\}\) is a set of \(n\) iid samples, then
    \[
    P_{T|\Theta}(D | \theta) = \prod_{i=1}^{d} \theta_i^{s_i} (1 - \theta_i)^{n-s_i}, \quad s_i = \sum_{j=1}^{n} x_i^{(j)}
    \]
Selecting priors

another example, problem 3.5.17 DHS (HW prob 3)

• in (b) you then show that if $\Theta$ is uniform (non-informative) the predictive distribution is

\[
P_{X|\Theta}(x | D) = \prod_{i=1}^{d} \left( \frac{S_i + 1}{n+2} \right)^{x_i} \left( 1 - \frac{S_i + 1}{n+2} \right)^{1-x_i}
\]

• in (d) you show that comparing with

\[
P_{X|\Theta}(x | \theta) = \prod_{i=1}^{d} \theta_i^{x_i} \left( 1 - \theta_i \right)^{1-x_i}
\]

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

  - **ML:** we know that
    
    $$\hat{\theta}_i = \frac{s_i}{n}$$

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

  - 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

\[
P_{T|\Theta}(D|\theta) = \prod_{i=1}^{d} \theta_{i}^{s_{i}} (1 - \theta_{i})^{n - s_{i}}, \quad s_{i} = \sum_{j=1}^{n} x_{i}^{(j)}
\]

- 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:
  \[
  \frac{s_i}{\theta_i} - \frac{(n-s_i)}{1-\theta_i} + \frac{1}{\theta_i} = 0 \iff \hat{\theta}_i = \frac{s_i + 1}{n+1}
  \]
- and
  \[
  P_{X|D}(x|D) = \prod_{i=1}^{d} \left( \frac{s_i + 1}{n+1} \right)^{x_i} \left( 1 - \frac{s_i + 1}{n+1} \right)^{1-x_i}
  \]
- 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:

\[
\frac{s_i}{\theta_i} - \frac{(n - s_i)}{1 - \theta_i} - \frac{1}{1 - \theta_i} = 0 \iff \hat{\theta}_i = \frac{s_i}{n+1}
\]

- and

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

- 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
  \[ P_{x|D}(x | D) = \prod_{i=1}^{d} \hat{\theta}_{i}^{x_{i}}(1 - \hat{\theta})^{1-x_{i}} \]

<table>
<thead>
<tr>
<th>Estimator</th>
<th>( \hat{\theta}_{i} )</th>
<th># tosses</th>
<th># “1”s</th>
<th>interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td>( s_{i}/n )</td>
<td>( n )</td>
<td>( s_{i} )</td>
<td></td>
</tr>
<tr>
<td>MAP non-informative</td>
<td>( s_{i}/n )</td>
<td>( n )</td>
<td>( s_{i} )</td>
<td>“the same”</td>
</tr>
<tr>
<td>MAP favor “1”s</td>
<td>( (s_{i}+1)/(n+1) )</td>
<td>( n+1 )</td>
<td>( s_{i}+1 )</td>
<td>“add one 1”</td>
</tr>
<tr>
<td>MAP favor “0”s</td>
<td>( s_{i}/(n+1) )</td>
<td>( n+1 )</td>
<td>( s_{i} )</td>
<td>“add one 0”</td>
</tr>
<tr>
<td>Bayes non-informative</td>
<td>( (s_{i}+1)/(n+2) )</td>
<td>( n+2 )</td>
<td>( s_{i}+1 )</td>
<td>“add one of each”</td>
</tr>
</tbody>
</table>

• 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 \( \frac{s_i}{n} \) is the “empty bin” problem
• for small \( n \), \( s_i \) is likely to be zero independently of the value of \( \theta_i \)
• 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 $C = (c_1, ..., c_w)$ 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

\[
d(C^i, C^j) = \sum_{k=1}^{w} p^i_k \log \left( \frac{p_k^i}{p_k^j} \right)
\]

- where the **probabilities** are the counts after normalization

\[
p^i_k = \frac{c^i_k}{\sum_k c^i_k}
\]

- problem: log goes to infinity when \( p^i_k = 0 \! \)
- 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 \( \pi_j \) is the probability of word \( j \).
- since the \( \pi_j \) 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

\[ P_{\pi}(\pi_1, \ldots, \pi_W) = \frac{\prod_{j=1}^{w} \pi_j^{u_j - 1}}{\prod_{k=1}^{w} \Gamma(\pi_j) \prod_{j=1}^{w} \Gamma(u_j)} \]

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

Prob 3 on HW

- on HW you will show that the posterior is

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

- 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?