Bayesian Statistics

In the previous lectures, especially on maximum Likelihood method, we focused on $P(D|\theta)$ but noted that, sometimes, when noise effects are large, it may be better not to take the max likelihood estimate. In other words, we can accept some amount of bias to reduce the impact of variance.

Our extreme example: x_i i.i.d $N(\mu, \sigma^2)$ $i = 1, \cdots, N$

MLE for μ: $\hat{\mu}_{MLE} = \bar{X}$

$$E\left(\bar{X} - \mu\right)^2 = \frac{\sigma^2}{N} \qquad MSE_{MLE} = \frac{\sigma^2}{N}$$

No bias, all variance.

However, if $|\mu| < \frac{\sigma}{\sqrt{N}}$ (Sample size too small)

$\hat{\mu} = 0$ has less error.

No variance, but bias.

$$MSE_0 = \mu^2 \qquad MSE_0 < MSE_{MLE}$$

Bayesian approach is one way of balancing bias-variance trade-off.

Now we allow talking about probability distribution of θ's themselves

$$\frac{P(\mathcal{D}|\theta)\, p_0(\theta)}{\int d\theta'\, P(\mathcal{D}|\theta')\, p_0(\theta')} = P(\theta|\mathcal{D})$$

($p_0(\theta)$: Prior; $P(\theta|\mathcal{D})$: Posterior; denominator: $P(\mathcal{D})$)

Note that $\log P(\theta|\mathcal{D})$

$$= \log P(\mathcal{D}|\theta) + \log p_0(\theta) - \log P(\mathcal{D})$$

- $\log P(\mathcal{D}|\theta)$: log likelihood
- $\log p_0(\theta)$: additional regularizer / penalty function for θ
- $-\log P(\mathcal{D})$: constant

If we want to find the "most likely" θ, we optimize $\log P(\theta|\mathcal{D})$. The additional regularizer / penalty alters $\hat{\theta}$ from $\hat{\theta}_{MLE}$.

This estimate is called the maximum a posteriori (MAP) estimate.

Several non-Bayesian approaches also advocate using additional penalties for estimation in order to control.

Bayesians (believe) there is more to Bayesian approach. Before we go there, let us look at some examples, though.

I will skip examples of discrete choice of hypotheses, as discussed in the beginning of Chapter 3. Let us start with the beta-binomial model.

$X_i \sim Ber(\theta)$ $\quad Prob[x_i=1]=\theta$, $Prob[x_i=0]=1-\theta$

X_i's independent.

[Ex: Head-tails for a possibly biased coin]

$$P(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0}$$

$$N_1 = \sum_{i=1}^{N} I(x_i = 1)$$

$$N_0 = \sum_{i=1}^{N} I(x_i = 0)$$

$$N_1 + N_0 = N$$

Note that N_1 is a sufficient statistics for θ. We don't need to know which order successes happened to infer θ.

$$N_1 \sim Bin(N, \theta)$$

$$Bin(k|N,\theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

Now let us choose a prior. Note that if we choose prior $p_0(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$, namely $\theta \sim Beta(a,b)$, then,

$$p(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)\, p_0(\theta)$$

$$\propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1}$$

So the posterior distribution follows a beta

distribution as well.

When the posterior and prior belongs to the same family of distributions the prior and the posterior are called conjugate distribution and the prior is called a conjugate prior for the likelihood function.

In this case, the prior is effectively adding some pseudo-counts to N_1, N_0.

The parameters a, b are called hyperparameters. We will later talk about how to choose these hyperparameters.

If we set $a=1, b=1$, we get an uniform prior.

In principle we could update things sequentially

prior $\xrightarrow{\text{data 1}}$ Posterior = new prior $\xrightarrow{\text{data 2}}$ new posterior

is equivalent to

prior $\xrightarrow{(\text{data 1}, \text{data 2})}$ new posterior

In many circumstances, when data arrives sequentially, you can have an online algorithm updating things sequentially.

Posterior mean and mode

Mode $\hat{\theta}_{map} = \arg\max_{\theta} \log\left[\theta^{N_1+a-1}(1-\theta)^{N_0+b-1}\right]$

θ deriv $\frac{N_1+a-1}{\theta} - \frac{N_0+b-1}{1-\theta} = 0$

$\Rightarrow \frac{\theta}{1-\theta} = \frac{N_1+a-1}{N_0+b-1} \Rightarrow \hat{\theta}_{MAP} = \frac{N_1+a-1}{N_0+N_1+a+b-2}$

$= \frac{N_1+a-1}{N+a+b-2}$

When $a=b=1$ $\hat{\theta}_{MAP} = \frac{N_1}{N} = \theta_{MLE}$

What is the average posterior θ?

$$\bar{\theta} = \frac{\int_0^1 d\theta\, \theta^{\gamma_1+1}(1-\theta)^{\gamma_2}}{\int_0^1 d\theta\, \theta^{\gamma_1}(1-\theta)^{\gamma_2}}$$

where $\gamma_1 = N_1 + a - 1 \qquad \gamma_2 = N_2 + b - 1$

$$\bar{\theta} = \frac{\int_0^1 d\theta\, \theta^{\gamma_1+1}(1-\theta)^{\gamma_2}}{\int_0^1 d\theta\, \theta^{\gamma_1}(1-\theta)^{\gamma_2}} = \frac{B(\gamma_1+2, \gamma_2+1)}{B(\gamma_1+1, \gamma_2+1)}$$

$$= \frac{\Gamma(\gamma_1+2)\,\Gamma(\gamma_2+1)/\Gamma(\gamma_1+\gamma_2+3)}{\Gamma(\gamma_1+1)\,\Gamma(\gamma_2+1)/\Gamma(\gamma_1+\gamma_2+2)}$$

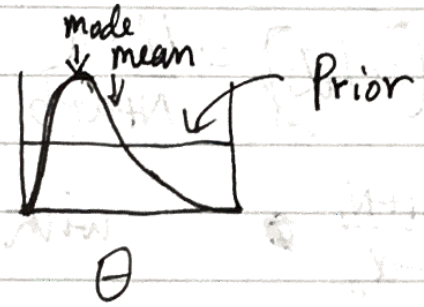
$$= \frac{\gamma_1+1}{\gamma_1+\gamma_2+2} = \frac{N_1+a}{N+a+b}$$

This is different from $\hat{\theta}_{MAP}$

Even if I took $a=b=1$,

I get $\bar{\theta} = \frac{N_1+1}{N+2}$

The mean posterior seems to have inserted 'pseudo counts' by itself. This is an example of Bayesian averaging 'regularizing' the estimate



Such ~~smoothing~~ 'smoothing' is important when data is sparse. Imagine $N_1 = 0$, no success observed so far. Can I conclude $\hat{\theta} = 0$? With uniform prior. $P(\theta|D) \propto (1-\theta)^N \approx e^{-N\theta}$

Intuitively that observation is saying that we believe $\theta \sim O(\frac{1}{N})$ and that is why perhaps we have not observed it.

$$\bar{\theta} = \frac{0+1}{N+2} = \frac{1}{N+2}$$

Posterior mean $\frac{N_1+1}{N+2}$ as θ estimate gives

Laplace's rule of succession and it is also known as add-one smoothing!

Note that $\bar{\theta} = E[\theta|\mathcal{D}] = \frac{N_1 + a}{N + a + b}$

$= \frac{\alpha_0 m_1 + N_1}{N + \alpha_0}$ with $\alpha_0 = a + b$

$m_1 = \frac{a}{a+b}$ ← Prior mean

$$E[\theta|\mathcal{D}] = \left(\frac{\alpha_0}{N+\alpha_0}\right) m_1 + \left(\frac{N}{N+\alpha_0}\right)\frac{N_1}{N}$$

m_1: Prior mean; $\frac{N_1}{N}$: MLE

$= \lambda m_1 + (1-\lambda)\hat{\theta}_{MLE}$

λ: Convex combination

As $N \to \infty$ $\bar{\theta}$ goes to $\hat{\theta}_{MLE}$

With lots of data, priors do not matter.

Posterior variance

$$\text{var}[\theta|\mathcal{D}] = E[\theta^2|\mathcal{D}] - E[\theta|\mathcal{D}]^2$$

$$= \frac{\Gamma(\gamma_1+3)\Gamma(\gamma_2+1)/\Gamma(\gamma_1+\gamma_2+4)}{\Gamma(\gamma_1+1)\Gamma(\gamma_2+1)/\Gamma(\gamma_1+\gamma_2+2)} - \left(\frac{\gamma_1+1}{\gamma_1+\gamma_2+2}\right)^2$$

$$= \frac{(\gamma_1+2)(\gamma_1+1)}{(\gamma_1+\gamma_2+3)(\gamma_1+\gamma_2+2)} - \frac{(\gamma_1+1)(\gamma_1+1)}{(\gamma_1+\gamma_2+2)(\gamma_1+\gamma_2+2)}$$

$$= \frac{\gamma_1+1}{\gamma_1+\gamma_2+2}\left[\frac{\gamma_1+2}{\gamma_1+\gamma_2+3} - \frac{\gamma_1+1}{\gamma_1+\gamma_2+2}\right]$$

$$= \frac{(\gamma_1+1)\left[(\gamma_1+\gamma_2)+4-\{\gamma_1+\gamma_2+3\}\right]}{(\gamma_1+\gamma_2+2)^2(\gamma_1+\gamma_2+3)}$$

$$= \frac{(\gamma_1+1)(\gamma_2+1)}{(\gamma_1+\gamma_2+2)^2(\gamma_1+\gamma_2+3)}$$

$$= \frac{(N_1+a)(N_0+b)}{(N_1+N_0+a+b)^2(N_1+N_0+a+b+1)}$$

When N is large compared to a, b

$$\text{Var}[\theta|\mathcal{D}] \simeq \frac{N_1 N_0}{N^3}$$

$$\simeq \frac{\hat{\theta}(1-\hat{\theta})}{N}$$

where $\hat{\theta} = \hat{\theta}_{MLE}$.

This is related to $\hat{\theta}_{MLE} = \frac{N_1}{N}$ following central limit theorem with the std/error bar

$$\sigma = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{N}}$$

Note that σ is the largest when $\hat{\theta} = \frac{1}{2}$, making is harder to decide whether a coin is fair, in comparison to establishing large bias.

More than two outcomes: Beta-Binomial
→ Dirichlet-multinomial

N 'dice' rolls $\{x_1, \cdots, x_N\}$ $x_i \in \{1, \cdots, K\}$

$$p(\mathcal{D}|\theta) = \prod_K \theta_k^{N_k} \qquad \sum_k \theta_k = 1 \qquad \sum_k N_k = N$$

For real dice $\theta_1, \cdots, \theta_6$

Honestly, if we were dealing with coin tosses or dice rolls, we can have N large enough so that all the Bayesian caution may not be too relevant unless $\theta \sim 1/N$. In practice one may be dealing with text classification or 'words' in DNA. For words in a natural language $K \sim 10^5$. For DNA sequences 6 bases long, $K = 4^6 \approx 4 \times 10^3$. Some words may be rare enough not to show up in the training data set, but appear in the test data. We could also have a very common word in the training data be missing in the test set. In all these cases, Bayesian average helps.

prior $\mathrm{Dir}(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_k \theta_k^{\alpha_k - 1}$ ($\alpha_1, \ldots, \alpha_K$)

Conjugate prior

Posterior $P(\theta|D) \propto P(D|\theta)\, p(\theta)$

$$\propto \prod_{k=1}^{K} \theta_k^{\alpha_k + N_k - 1}$$

$$= \mathrm{Dir}(\theta | \alpha_1 + N_1, \ldots, \alpha_K + N_K)$$

Optimize subject to constraint $\sum_k \theta_k = 1$

Mode $l(\theta, \lambda) = \sum_k (N_k + \alpha_k - 1) \log \theta_k + \lambda \left(1 - \sum_k \theta_k\right)$

$\frac{\partial l}{\partial \theta_k} = 0 \rightarrow \frac{N_k + \alpha_k - 1}{\theta_k} = \lambda$ or $(N_k + \alpha_k - 1) = \lambda \theta_k$

Sum over k $\quad N + \alpha_0 - K = \lambda$

$\alpha_0 = \sum_{k=1}^{K} \alpha_k$

So $\hat{\theta}_k = \frac{N_k + \alpha_k - 1}{\lambda} = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K}$

That is the MAP estimate $\hat{\theta}_k = \frac{N_k}{N}$ for MLE.

Posterior predictive

$$p(X=j|\mathcal{D}) = \int p(x=j|\theta)\, p(\theta|\mathcal{D})\, d\theta$$

$$= \int p(x=j|\theta_j)\, P(\theta|\mathcal{D})\, d\theta$$

$$= \int \theta_j\, p(\theta_j|\mathcal{D})\, d\theta_j = \bar{\theta}_j$$

$$p(\theta_j|\mathcal{D}) = \frac{\int \theta_j^{\gamma_j} \prod_{i \neq j} \theta_i^{\gamma_i} \, d\theta_{-j}}{\int \prod_i \theta_i^{\gamma_i} \, d\theta}$$

$$= \frac{\theta_j^{\gamma_j}(1-\theta_j)^{\sum_{i \neq j}\gamma_i + K-1}}{\int d\theta_j\, \theta_j^{\gamma_j}(1-\theta_j)^{\sum \gamma_i + K-1}}$$

$$\gamma_j = N_j + \alpha_j - 1$$

$$\bar{\theta}_j = \frac{\gamma_j + 1}{\gamma_j + \sum_{i \neq j}\gamma_i + K - 1 + 1} = \frac{N_j + \alpha_j}{N + \alpha_0}$$

Once more the porterior predictive avoids the zero count problem [Even when we have a flat prior; $\alpha_j = 1$].

Bag of word model, document classification
Naive Bayes.

Bag of word model

Mary had a little lamb, little lamb, little lamb,
Mary had a little lamb, its fleece as white as snow

Drop 'stop words'

Vocab list

mary	lamb	little	big	fleece	white	black	snow
1	2	3	4	5	6	7	8

rain unk.
0 10↑ unknown

mary unk little lamb little lamb little lamb
1 10 3 2 3 2 3 2
mary unk little lamb unk fleece white snow
1 10 3 2 10 5 6 8

$N = 16$ $\alpha_j = 1$

$$\bar{\theta}_j = \frac{N_j + 1}{16 + 10}$$

$\bar{\theta}_{mary} = \frac{2+1}{16+10} = \frac{3}{26}$, $\bar{\theta}_{big} = \frac{1}{16+10} = \frac{1}{26}$

big never shows up

Naive Bayes! $(x_1, \ldots, x_D)$ [D features]

'Naive' assumption feature conditionally independe
(conditioned upon class).

$$p(x|y=c,\theta) = \prod_{j=1}^{D} p(x_j|y=c,\theta_{jc})$$

Example: 1) For real-valued [redacted] features $p(x|c,\theta) = \prod_{j=1}^{D} N(x_j|\mu_{jc}, \sigma_{jc}^2$

2) Binary features $x_j \in \{0,1\}$, $p(x|c,\theta) = \prod_{j=1}^{D} \text{Ber}(x_j|\theta$

Let us see an application to document classification:

Bernoulli document model:

D 'feature' words

x_1	x_2	-	x_{D-1}	x_D
0	1		1	0
↑	↑		↑	↑
word 1 absent	word 2 present	..	word D-1 present	word D absent

} We throw away the the details of word abundance

Bayesian priors $\vec{\pi} = (\pi_1, \ldots \pi_C)$ → Probs of belonging to different clas

$p(\vec{\pi}) \rightarrow \text{Dir}(\alpha)$

$$p(\theta) = \prod_{j=1}^{D} \prod_{c=1}^{C} p(\theta_{jc}) \rightarrow \text{Beta}(\beta_0, \beta_1)$$

With training data $\mathcal{D} \rightarrow N_1, \ldots, N_C$ (numbers in different classes)

N_{jc} = No. of word j in docs of class c

$$p(\theta|\mathcal{D}) = p(\pi|\mathcal{D}) \prod_{j=1}^{D} \prod_{c=1}^{C} p(\theta_{jc}|\mathcal{D})$$

$$p(\pi|\mathcal{D}) = \mathrm{Dir}(N_1+\alpha_1, \ldots, N_C+\alpha_C) \propto \pi_1^{N_1+\alpha_1} \cdots \pi_C^{N_C+\alpha_C}$$

$$p(\theta_{jc}|\mathcal{D}) = \mathrm{Beta}(N_c - N_{jc} + \beta_0, N_{jc}+\beta_1) \propto \theta_{jc}^{N_{jc}+\beta_1} (1-\theta_{jc})^{N_c - N_{jc} + \beta_0}$$

For posterior predictive, integrate out parameters

$$p(y=c|x,\mathcal{D}) \propto \int \mathrm{Cat}(y=c|\pi)\, p(\pi|\mathcal{D})\, d\pi \prod_{j=1}^{D} \left[\int \mathrm{Ber}(x_j|y=c,\theta_{jc})\, p(\theta_{jc}|\mathcal{D})\, d\theta_{jc} \right]$$

$$= \bar{\pi}_c \prod_{j=1}^{D} \bar{\theta}_{jc}^{\mathbb{I}(x_j=1)} (1-\bar{\theta}_{jc})^{\mathbb{I}(x_j=0)}$$

with $\bar{\pi}_c = \frac{N_c+\alpha_c}{N+\alpha_0}$ $\qquad \bar{\theta}_{jc} = \frac{N_{jc}+\beta_1}{N_c+\beta_0+\beta_1}$

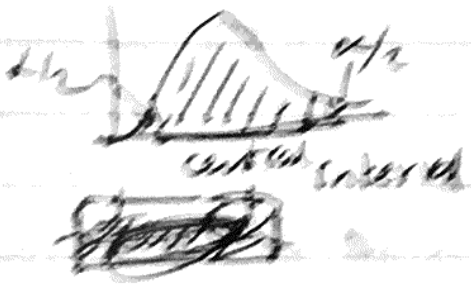
$\alpha_0 = \sum_c \alpha_c$

Note that $\bar{\theta}_{jc}$ is being used for evaluating the class rather than $(\hat{\theta}_{jc})_{MAP}$

One can use the multinomial model using word frequency, but it could be problematic if certain words are 'bursty': they appear many times when they do in a document, but does not at all appear in most documents.

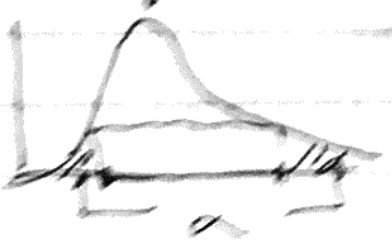
Back to general issues with Bayesian method we should take advantage of the fact that Bayesian approach gives us a whole distribution of θ.

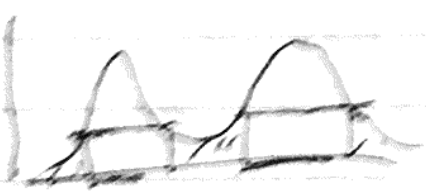
We could do confidence intervals → credible intervals



If the distribution is skewed

highest posterior density region





If the posterior distribution is not analytically tractable, one can use Monte Carlo to sample from it and get p-values.

Another important aspect of Bayesian approach is its take on model selection.

Bayesian model selection

Remember the polynomial fit exercise where choosing different degrees led to models with different complexities. How does Baysian approach choose among them?

As opposed to the cross validation approach, which required doing many fits, we compute, for a model m

$$P(m|\mathcal{D}) = \frac{P(\mathcal{D}|m)\,p(m)}{\sum_{m'} P(\mathcal{D}|m')\,p(m')}$$

Biggest $P(m|\mathcal{D})$ wins.

For uniform prior ($p(m)$ constant) we just take largest $P(\mathcal{D}|m)$

$$P(\mathcal{D}|m) = \int P(\mathcal{D}|\theta)\,P(\theta|m)\,d\theta$$

integrate over all choices of parameters

Evidence,
Marginal likelihood,
Integrated likelihood

Bayesian Occam' razor

Occam' razor: One should pick the simplest model that adequately explains the data.

Use $P(\mathcal{D} \mid \hat{\theta}_m)$ to choose between models

Two complex models

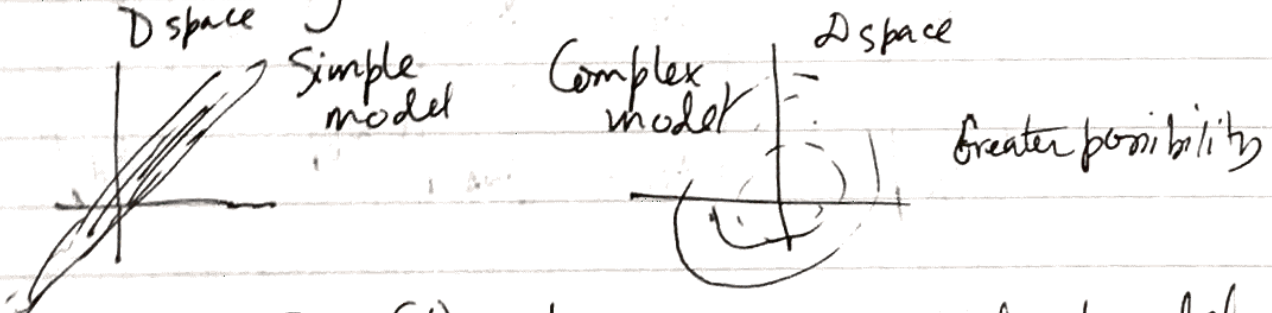
$$P(\mathcal{D}) = P(y_1)\, P(y_2 \mid y_1)\, P(y_3 \mid y_1, y_2) \cdots P(y_N \mid y_{1:N-1})$$

$P(y_3 \mid y_1, y_2)$: good likelihood by overfitting initial data

$P(y_N \mid y_{1:N-1})$: Bad likelihood from later pieces.

Like training good, validation bad.

Another way is to see it in that



Since $\sum_{D'} P(D') = 1$

more spread out model usually produces smaller $P(D)$

Example $x_i \sim N(\mu_i, \sigma^2)$

$P(\mu) \propto e^{-\mu^2/2\sigma_m^2}$

Model 1 $\mu_1 = \mu_2$

Model 2 μ_1 μ_2 independent