Bayesian Model Selection

Last time: bias & variance as a way to understand how model complexity affects generalization.

(freq: data are random $\mathbb{E}_{\mathcal{D}}[\hat{f}_{\mathcal{D}}(x)]$

Today: Bayesian view: data are fixed; params are random

$\downarrow$

is a formula for many things

1) compute $p(w|X,Y)$ } posterior over params

2) compute $p(y^*|x^*,X,Y) = \int_w p(y^*|x^*,w)\,p(w|X,Y)$

prediction — some new point

posterior predictive }

(BAYES is all about model-averaging)

3) compute $p(Y|X) = \underline{\int_W p(Y|X,w)\,p(w)}$ } marginal likelihood

Posterior over params:

assume prior over w doesn't dep. on x

Bayes rule: $p(w|X,Y) = \dfrac{P(Y|X,w)\,P(W|X)}{P(Y|X)}$ ← notice: this is a constant wrt w.

$\propto P(Y|X,W)\,P(W)$

$\propto P(W,Y|X)$

depending on $P(Y|X,W)$, $P(W)$ → easy or very hard

↓

conjugate.

Example w/ Beta-Bernoulli Model

Model: coin come up "1" with prob. θ } x is the flip

$p(x|\theta) = \theta^x(1-\theta)^{1-x}$ likelihood

$p(x_1 \dots x_N|\theta) = \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{1-x_n}$

$= \theta^{n_1}(1-\theta)^{n_0}$, where $n_1+n_0=N$

Aside: from this, we could compute θ_{MLE}, θ Maximum likelihood,

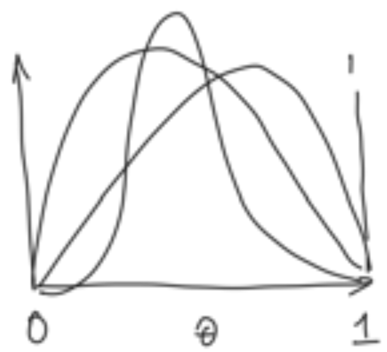
$\max_\theta \log p(x_1 \dots x_n|\theta) = \max_\theta n_1 \log\theta + n_0 \log(1-\theta)$

$\Rightarrow \theta_{MLE} = \dfrac{n_1}{n_0+n_1}$

Now, let's put a prior on θ. Beta distribution: $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\, \mathbb{I}_{(0,1)}$

for $\alpha, \beta > 0$

(θ in $[0,1]$)



for $\alpha, \beta > 1$

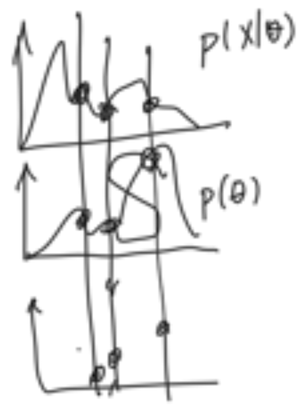
if $\alpha, \beta = 1 \Rightarrow$ uniform

if $\alpha, \beta < 1$
"sparsity favoring"

Why this choice?

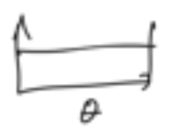
$$p(\theta \mid X) \propto \underline{p(X \mid \theta)}\ \underline{p(\theta)}$$
$$\propto \underline{\theta^{n_1}(1-\theta)^{n_0}}\ \underline{\theta^{\alpha-1}(1-\theta)^{\beta-1}}$$
$$= \underline{\theta^{n_1+\alpha-1}(1-\theta)^{n_0+\beta-1}} \quad \text{looks like another Beta!!}$$

$$\underset{p(\theta)}{\mathrm{Beta}(\alpha,\beta)} \longrightarrow \underset{p(\theta \mid X)}{\boxed{\mathrm{Beta}(n_1+\alpha,\ n_0+\beta)}}$$



Fill in some data:

- $\alpha = 1$, $\beta = 1$
- we see 2 heads: $n_0 = 0$, $n_1 = 2$
- new $\underline{\mathrm{Beta}(3, 1)}$



With this dist, we can compute M.A.P. (last time)

$$\max_\theta \log p(X \mid \theta)\, p(\theta) \Rightarrow \boxed{\frac{n_1+\alpha-1}{n_1+n_0+\alpha+\beta-2}} = \theta_{MAP} \quad \text{mode of Beta}$$

with this dist, we can $\underline{\text{also}}$ compute $\underline{\text{posterior predictive}}$

$$p(x=1 \mid x_1 \cdots x_N) = \int_\theta \underbrace{p(x=1 \mid \theta)}_{\text{predict}}\ \underbrace{p(\theta \mid x_1 \cdots x_N)}_{\text{posterior}}$$
$$= \int \theta\, p(\theta \mid x_1 \cdots x_N) = E_{p(\theta \mid x_1 \cdots x_N)}[\theta] = \underbrace{\frac{\alpha+n_1}{\alpha+\beta+n_1+n_0}}_{\text{mean of Beta}}$$

$\theta_{\ } = \frac{n_1}{\ } = 1 \quad \operatorname{argmax} p(x \mid \theta)$

$\theta_{MLE} = \frac{n_1}{n_1+n_0} = 1$ argmax ...

$\theta_{MAP} = 1$ argmax $p(x|\theta)\,p(\theta)$

$\theta_{PP} = 3/4$ $E_{p(\theta|x_1 \dots x_N)}[\theta]$

in a diff problem: we could compare

$p(y|x, \theta_{MLE})$

$p(y|x, \theta_{MAP})$

$p(y|x, X, Y) \leftarrow PP$

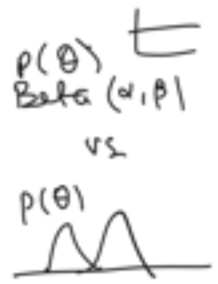
finally: $\underbrace{p(x_1 \dots x_N)}_{p(X)} = \int \underbrace{\theta^{n_1}(1-\theta)^{n_0}}_{p(X|\theta)} \underbrace{\theta^{\alpha-1}(1-\theta)^{\beta-1} Z_{\alpha,\beta}}_{p(\theta)}$

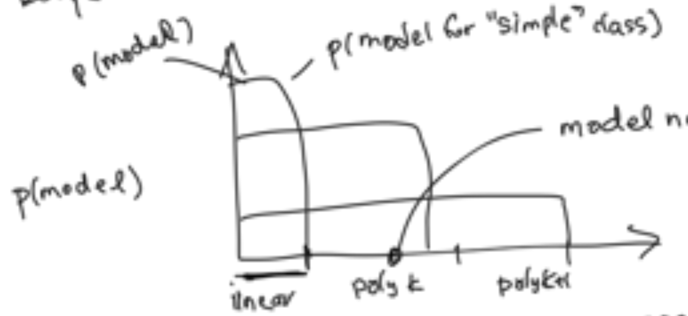
looks like $Beta(\alpha+n_1, \beta+n_0)$ $\frac{Z_{\alpha+n_1,\beta+n_0}}{Z_{\alpha+n_1,\beta+n_0}}$

$= \underbrace{\frac{Z_{\alpha,\beta}}{Z_{\alpha+n_1,\beta+n_0}}}_{\text{marginal likelihood.}} \int \cancel{Beta(\alpha+n_1, \beta+n_0)} \rightarrow 1$

Why is this marginal likelihood useful?

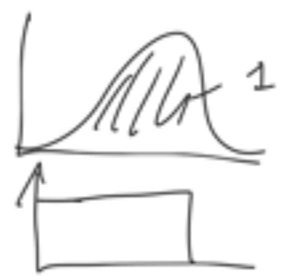
Bayesian Occam's Razor: $p(X)$ — implicitly $p(X|\text{model class})$



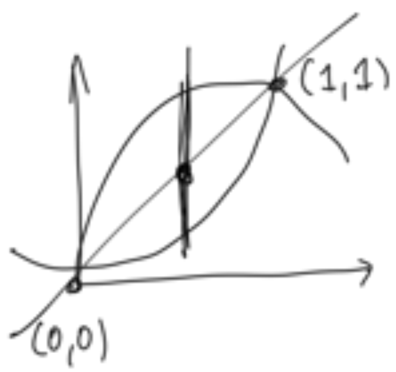


models: less to more complex →

(linear vs. poly k vs. poly w/ bigger k)



Concept Check



$y = f(x) + \underbrace{\varepsilon}_{\text{assume tiny}}$

1. What parabolas can we fit to these data, assume noise is very small. (what is implication of small noise)

→ parabola will go through data

2. $y = a_0 + a_1 x + a_2 x^2$ can we define a single parabola that fits data?

$a_0 = 0$

$a_2 = 1 - a_1$

$1 = a_1 + a_2$

3. improper prior on a_2 [$p(a_2) = 1$ for all a_2], and unif $[-1, 1]$ for a_1.
   What is posterior predictive at $x = 1/2$.

   if $a_1 = 1$: $a_2 = 0 \rightarrow 0$
   if $a_1 = -1$: $a_2 = 2 \rightarrow 1/2$

   $y(x=\frac{1}{2}) = \frac{1}{4} + \frac{1}{4}a_1 \Rightarrow$ avg $\frac{1}{4}$

4. What would be different if we assumed a linear model $y = a_0 + a_1 x$?
   (marginal likelihood)

   linear model: <u>only one model</u>

   posterior pred: $1/2$

   marg lik: