Bayesian Model Averaging

Model <u>Selection</u>: pick one of several models $M_1, M_2, \ldots M_K$

Model <u>Averaging</u>: don't pick one model but combine predictions from several models.

Suppose Δ is a quantity you wish to predict. Examples:
- effect size
- utility of a decision
- statistical parameter

Each model M_k gives a predicted value of Δ.

→ better predictions of Δ using <u>weighted average</u> of model's prediction of Δ. (weighted using each model's <u>posterior probability</u>).

→ In other words, we wish to compute $E[\Delta|X]$ for our data X using a sum over our models.

Conditional Expectation: $E[\Delta|X] = \int \Delta \Pr(\Delta|X) d\Delta$ (assuming Δ is cont.) (1)

where $\Pr(\Delta|X)$ can be written as a sum over models:

$$\Pr(\Delta|X) = \sum_{k=1}^{K} \Pr(\Delta, M_k|X) = \sum_{k=1}^{K} \Pr(\Delta|M_k, X)\Pr(M_k|X)$$

where does this come from?

$\Pr(A,B) = \Pr(A|B)\Pr(B)$ → now condition both sides

$P(A,B|C) = \Pr(A|B,C)\Pr(B|C)$ → compare

$$\underbrace{\Pr(\Delta|X)}_{\text{this prob. is the average}} = \sum_k \underbrace{\Pr(\Delta|M_k, X)}_{\text{prob } \Delta \text{ under each model } k} \underbrace{\Pr(M_k|X)}_{\text{weighted by posterior prob of that model}} \qquad (2)$$

Posterior Prob. of model M_k:

$$\Pr(M_k|X) = \frac{\Pr(X|M_k)\Pr(M_k)}{\sum_{\ell=1}^{K} \Pr(X|M_\ell)\Pr(M_\ell)} \qquad (3)$$

(from Bayes Thm)
(denom. is $\Pr(X)$)

and ↘

$$Pr(X|M_k) = \int Pr(X|\theta_k, M_k)\, Pr(\theta_k|M_k)\, d\theta_k$$

is the likelihood of model M_k w/ parameter(s) θ_k and prior prob. $Pr(\theta_k|M_k)$ for those parameters under model M_k.

→ If M_k is a normal distribution, then $\theta_k = (\mu_k, \sigma_k^2)$.

With these equations we can now compute $E[\Delta|X]$: ✱

$$E[\Delta|X] = \int \Delta\, Pr(\Delta|X)\, d\Delta \qquad \leftarrow \text{Plug Eq 2 into Eq 1}$$

$$= \int \Delta \left(\sum_k Pr(\Delta|M_k, X)\, Pr(M_k|X)\right) d\Delta \qquad \text{Exchange } \Sigma \text{ and } \int$$

$$= \sum_k \left(\int \Delta\, Pr(\Delta|M_k, X)\, d\Delta\right) Pr(M_k|X) \qquad \text{posterior prob (Eq 3) does not depend on } \Delta$$

$$= \sum_k \hat{\Delta}_k\, Pr(M_k|X) \qquad \text{where } \hat{\Delta}_k = \int \Delta\, Pr(\Delta|M_k, X)\, d\Delta = E[\Delta|M_k, X]$$

Now we can use $E[\Delta|X]$ as a better overall prediction of Δ, but there are some challenges:

1. Where do models $M_1, \ldots M_K$ come from? How to specify their prior probs. $Pr(M_k)$ in Eq 3?

2. Computational challenges:

   - Number of terms summed over in Eq 2 can be enormous
   - Integrals w/in Eq. 2 $(Pr(X|M_k))$ can be difficult to compute.

We will learn some techniques to address some of these challenges

<u>Remark</u>: Model averaging in the context of machine learning is a type of <u>ensemble learning</u> ⇒ very powerful!

✱ It is also important to compute <u>variance</u> in Δ, $Var(\Delta|X)$ → see supp.

Bayesian Model Averaging supplement — variance of Δ

Knowing $E[\Delta|X]$ benefits from also knowing $Var(\Delta|X)$: how <u>reliable</u> are our predictions?

Recall: $Var(A) = E\left[(A - E[A])^2\right] = E[A^2] - (E[A])^2$ for R.V. A

Likewise, for the conditional variance:

$$Var(A|B) = E\left[(A - E[A|B])^2 \,\middle|\, B\right]$$
$$= E[A^2|B] - (E[A|B])^2$$

To compute $Var(\Delta|X)$ requires knowing $E[\Delta|X]$, which we have, and $E[\Delta^2|X]$. Let's focus on the latter.

$$E[\Delta^2|X] = \int \Delta^2 \Pr(\Delta|X)\, d\Delta$$

proceed as we did w/ $E[\Delta|X]$

$$= \int \Delta^2 \sum_k \Pr(\Delta|M_k, X) \Pr(M_k|X)\, d\Delta$$
$$= \sum_k \underbrace{\left(\int \Delta^2 \Pr(\Delta|M_k, X)\, d\Delta\right)}_{\text{write this in terms of a variance}} \Pr(M_k|X)$$

$$\overbrace{\int \Delta^2 \Pr(\Delta|M_k, X)\, d\Delta}^{} - \Big(\underbrace{\int \Delta \Pr(\Delta|M_k, X)\, d\Delta}_{\text{recognize that this is } (\hat{\Delta}_k)^2}\Big)^2 = Var(\Delta|M_k X)$$

$\downarrow$ plug back in

$$= \sum_k \left(Var(\Delta|M_k X) + \hat{\Delta}_k^2\right) \Pr(M_k|X)$$ and plug this back into $Var(\Delta|X)$

$$Var(\Delta|X) = E[\Delta^2|X] - (E[\Delta|X])^2$$
$$= \sum_k \left(Var(\Delta|M_k, X) + \hat{\Delta}_k^2\right) \Pr(M_k|X) - (E[\Delta|X])^2, \quad \hat{\Delta}_k^2 = E[\Delta|M_k X]$$