Probability Theory

Kevin P. Murphy: Probabilistic Machine Learning Ch4

Presented by Wending Liu

RSE ML and Econometrics Reading Group

Oct 30 2023

Table of Contents

Why MLE?

- We use **Kullback Leibler divergence** to measure the distance between $p(\boldsymbol{y} \mid \theta)$ and the empirical distribution $p_{\mathcal{D}}(\boldsymbol{y}) \triangleq \frac{1}{N} \sum_{n=1}^{N} \delta(\boldsymbol{y} - \boldsymbol{y}_n)$.

$$\mathrm{KL}(p_{\mathcal{D}} \| p(\boldsymbol{y} \mid \theta)) = \sum_{\boldsymbol{y}} [p_{\mathcal{D}}(\boldsymbol{y}) \log p_{\mathcal{D}}(\boldsymbol{y}) - p_{\mathcal{D}}(\boldsymbol{y}) \log p(\boldsymbol{y} \mid \theta)]$$

$$= -\mathbb{H}(p_{\mathcal{D}}) - \frac{1}{N} \sum_{n=1}^{N} \log p(\boldsymbol{y}_n \mid \boldsymbol{\theta})$$

- Unconditional (unsupervised) model:

$$\hat{\boldsymbol{\theta}}_{\mathrm{mle}} = \operatorname*{argmin}_{\boldsymbol{\theta}} - \sum_{n=1}^{N} \log p(\boldsymbol{y}_n \mid \boldsymbol{\theta})$$

- Conditional (supervised) model:

$$\hat{\boldsymbol{\theta}}_{\mathrm{mle}} = \operatorname*{argmin}_{\boldsymbol{\theta}} - \sum_{n=1}^{N} \log p(\boldsymbol{y}_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})$$

Example: MLE for the Bernoulli distribution

- Suppose Y is a random variable representing a coin toss. Let $\theta = p(Y = 1)$ be the probability of heads.

$$\begin{aligned}
\text{NLL}(\theta) &= -\log \prod_{n=1}^{N} p\left(y_n \mid \theta\right) \\
&= -\log \prod_{n=1}^{N} \theta^{\mathbb{I}(y_n=1)}(1-\theta)^{\mathbb{I}(y_n=0)} \\
&= -\sum_{n=1}^{N} \left[\mathbb{I}\left(y_n = 1\right) \log \theta + \mathbb{I}\left(y_n = 0\right) \log(1-\theta)\right] \\
&= -\left[N_1 \log \theta + N_0 \log(1-\theta)\right]
\end{aligned}$$

- $N_1 = \sum_{n=1}^{N} \mathbb{I}\left(y_n = 1\right)$ and $N_0 = \sum_{n=1}^{N} \mathbb{I}\left(y_n = 0\right)$, are called the **sufficient statistics of the data**.
- $\hat{\theta}_{\text{mle}} = \frac{N_1}{N_0 + N_1}$.

Example: MLE for linear regression

- Conditional likelihood: $p(y \mid \boldsymbol{x}; \boldsymbol{\theta}) = \mathcal{N}\left(y \mid \boldsymbol{w}^T \boldsymbol{x}, \sigma^2\right)$.
- $\mathrm{NLL}(\boldsymbol{w}) = -\sum_{n=1}^{N} \log \left[\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n\right)^2\right)\right]$
- Let's assume σ^2 is known, and define **residual sum of squares**:

$$\mathrm{RSS}(\boldsymbol{w}) = \sum_{n=1}^{N}\left(y_n - \boldsymbol{w}^\top \boldsymbol{x}_n\right)^2 = \|\mathbf{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 = (\mathbf{X}\boldsymbol{w} - \boldsymbol{y})^\top(\mathbf{X}\boldsymbol{w} - \boldsymbol{y})$$

- $$\hat{\boldsymbol{w}}_{\mathrm{mle}} \triangleq \operatorname*{argmin}_{\boldsymbol{w}} \mathrm{RSS}(\boldsymbol{w}) = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \boldsymbol{y}$$

- OLS estimate is equivalent to the MLE for the linear regression model with Gaussian likelihood.

Empirical risk minimization (ERM)

- We can generalize MLE by replacing the (conditional) log loss with any other loss function, to get

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^{N} \ell\left(\boldsymbol{y}_n, \boldsymbol{\theta}; \boldsymbol{x}_n\right)$$

  This is known as **empirical risk minimization** or ERM.

- 0-1 loss:

$$\ell_{01}\left(\boldsymbol{y}_n, \boldsymbol{\theta}; \boldsymbol{x}_n\right) = \begin{cases} 0 & \text{if } \boldsymbol{y}_n = f\left(\boldsymbol{x}_n; \boldsymbol{\theta}\right) \\ 1 & \text{if } \boldsymbol{y}_n \neq f\left(\boldsymbol{x}_n; \boldsymbol{\theta}\right) \end{cases}$$

- We can smooth the 0-1 loss by using the surrogate loss:

$$\ell_{ll}(\tilde{y}, \eta) = \log\left(1 + e^{-\tilde{y}\eta}\right).$$

  where true label $\tilde{y} \in \{-1, 1\}$, $\eta = f(\boldsymbol{x}; \boldsymbol{\theta})$ is the log odds.

The method of moments

- Computing the MLE requires solving the equation $\nabla_\theta \text{NLL}(\theta) = \mathbf{0}$.
- A simpler approach is **method of moments** (MOM), which finds the estimates by equating the empirical moments to the theoretical moments.
- The kth order theoretical moments is given by $\mu_k = \mathbb{E}\left[Y^k\right]$, and the empirical moments are given by $\hat{\mu}_k = \frac{1}{N}\sum_{n=1}^{n} y_n^k$.
- MOM is theoretically inferior to the MLE approach, since it may not use all the data as efficiently.
- MOM can sometimes produce inconsistent results.
- However, when it produces valid estimates, it can be used to initialize iterative algorithms that are used to optimize the NLL, thus combining the **computational efficiency of MOM** with the **statistical accuracy of MLE**.

Example: MOM v.s. MLE for the uniform distribution

- $Y \sim \text{Unif}(\theta_1, \theta_2)$. $p(y \mid \theta) = \frac{1}{\theta_2 - \theta_1} \mathbb{I}(\theta_1 \leq y \leq \theta_2)$.
- The first two moments are

$$\mu_1 = \mathbb{E}[Y] = \frac{1}{2}(\theta_1 + \theta_2)$$

$$\mu_2 = \mathbb{E}\left[Y^2\right] = \frac{1}{3}\left(\theta_1^2 + \theta_1\theta_2 + \theta_2^2\right).$$

- $\mathcal{D} = \{0, 0, 0, 0, 1\}$.
- $\hat{\theta}_1^{mom} \approx -0.493$, $\hat{\theta}_2^{mom} \approx 0.893$.
- $p(\mathcal{D} \mid \boldsymbol{\theta}) = (\theta_2 - \theta_1)^{-N} \mathbb{I}(y_{min} \geq \theta_1) \mathbb{I}(y_{max} \leq \theta_2)$.
- $\hat{\theta}_1^{mle} = 0, \hat{\theta}_2^{mle} = 1$.

Table of Contents

Overfitting, Regularization and MAP estimation

- Suppose we flip a coin three times and observe three heads.
- $\hat{\theta}_{mle} = 3/(3+0) = 1$.
- MLE is prone to overfitting.
- The main solution to overfitting is to use regularization:

$$\mathcal{L}(\boldsymbol{\theta}; \lambda) = \left[-\sum_{n=1}^{N} \ell\left(\boldsymbol{y}_n, \boldsymbol{\theta}; \boldsymbol{x}_n\right) \right] + \lambda C(\boldsymbol{\theta})$$

  Where $\lambda \geq 0$ is the **regularization parameter**, and $C(\theta)$ is some form of **complexity penalty**.

- $\lambda = 1, C(\boldsymbol{\theta}) = -\log p(\theta) \rightarrow$ **MAP** (maximum a posterior estimation):

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p(\boldsymbol{\theta} \mid \mathcal{D}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} [\log p(\mathcal{D} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \text{const}]$$

Example: MAP estimation for the Bernoulli distribution

- Consider again the coin tossing example.
- Now suppose we have a prior

$$p(\theta) = \text{Beta}(\theta \mid a, b) = \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1}.$$

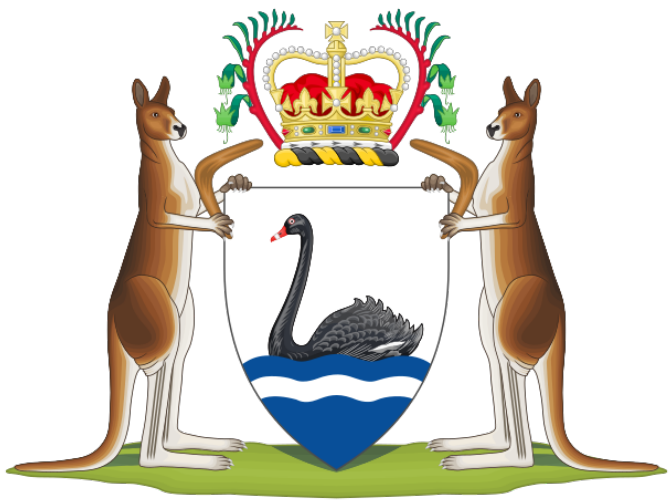
- Log likelihood + log prior:

$$\begin{aligned} &\log p(\mathcal{D} \mid \theta) + \log p(\theta) \\ &= [N_1 \log \theta + N_0 \log(1-\theta)] + [(a-1)\log\theta + (b-1)\log(1-\theta)]. \end{aligned}$$

- $\hat{\theta}_{\text{map}} = \frac{N_1+a-1}{N_0+N_1+a+b-2}$.
- If $a = b = 2$, then $\hat{\theta}_{\text{map}} = \frac{N_1+1}{N_0+N_1+2}$. (add one smoothing)

Black swan paradox

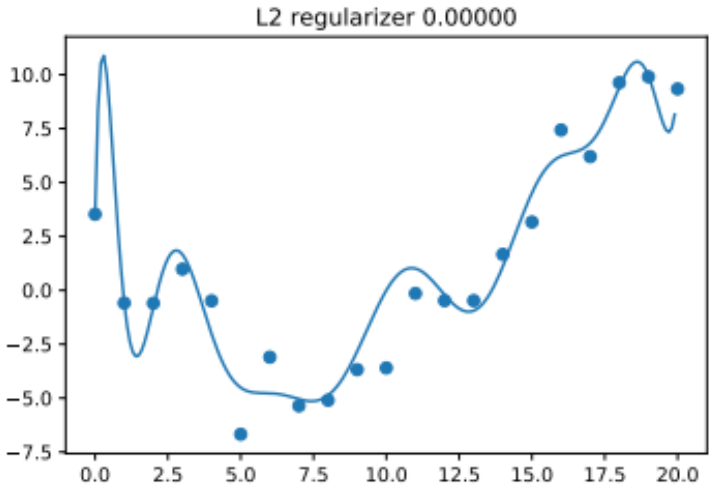
- "All swans are white" is an ancient Western conception.
- A black swan was a metaphor for something that could not exist.
- But black swan was discovered in Australia in 1697.
- This paradox was used to illustrate **the problem of induction**.
- The solution to the paradox is to admit that induction is in general impossible, and that the best we can do is to make plausible guesses about what the future might hold, by combining the empirical data with prior knowledge.



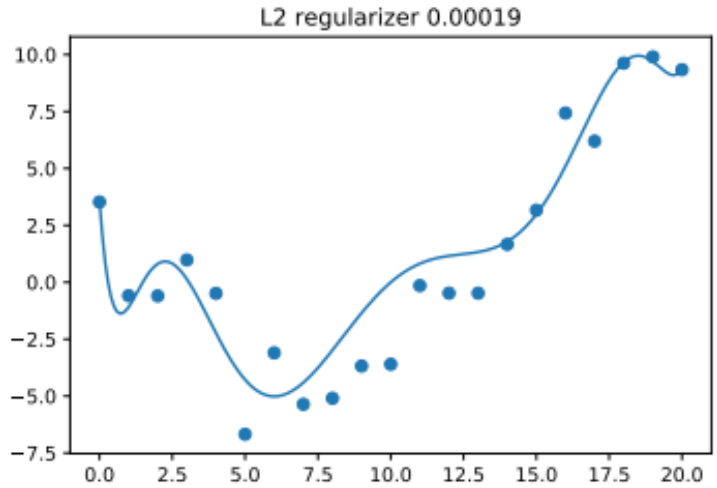
The Coat of Arms of Western Australia

Example: weight decay and ridge regression

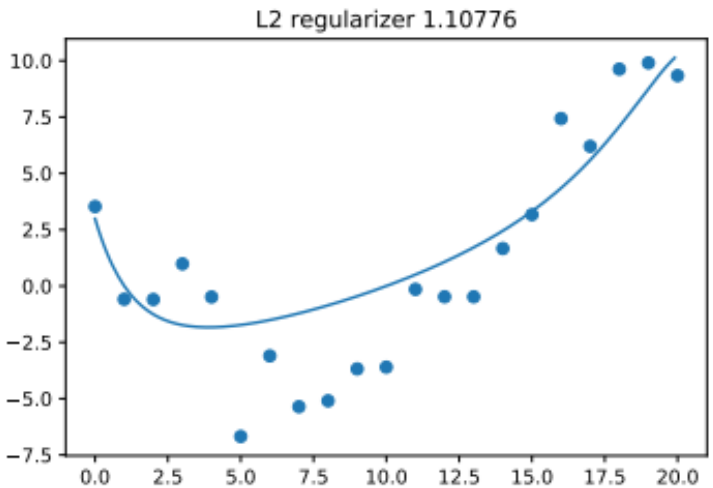
- Consider using a $D = 14$ polynomial to fit the $N = 21$ data curve.
- predicator: $f(x; w) = \mathbf{w} \cdot [1, x, x^2, \ldots, x^D]$.
- We use a Gaussian prior $p(w) = N(0, \frac{\lambda^{-1}}{2}\mathbb{I})$ to penalize large coefficients.
- $\hat{\boldsymbol{w}}_{\text{map}} = \underset{\boldsymbol{w}}{\text{argmin}}\, \text{NLL}(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_2^2$. (ridge regression)
- λ too small $\rightarrow$ **overfitting**; λ too large $\rightarrow$ **underfitting**.


L2 regularizer 0.00000

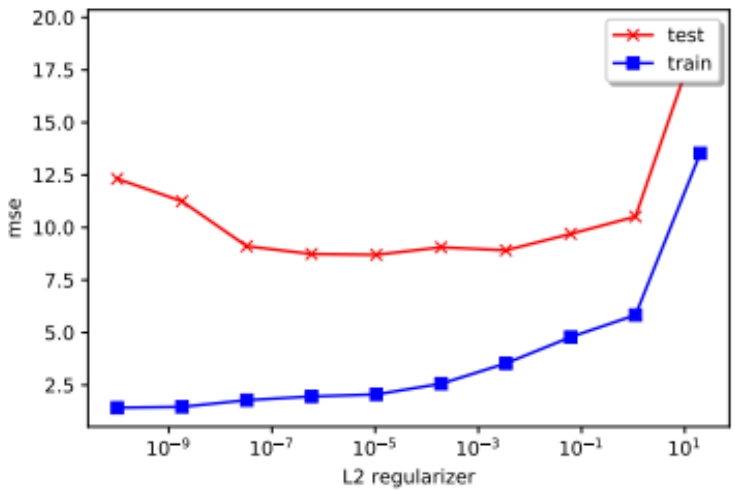
(a)


L2 regularizer 0.00019

(b)


L2 regularizer 1.10776

(c)


test
train
mse
L2 regularizer

(d)

Cross-Validation

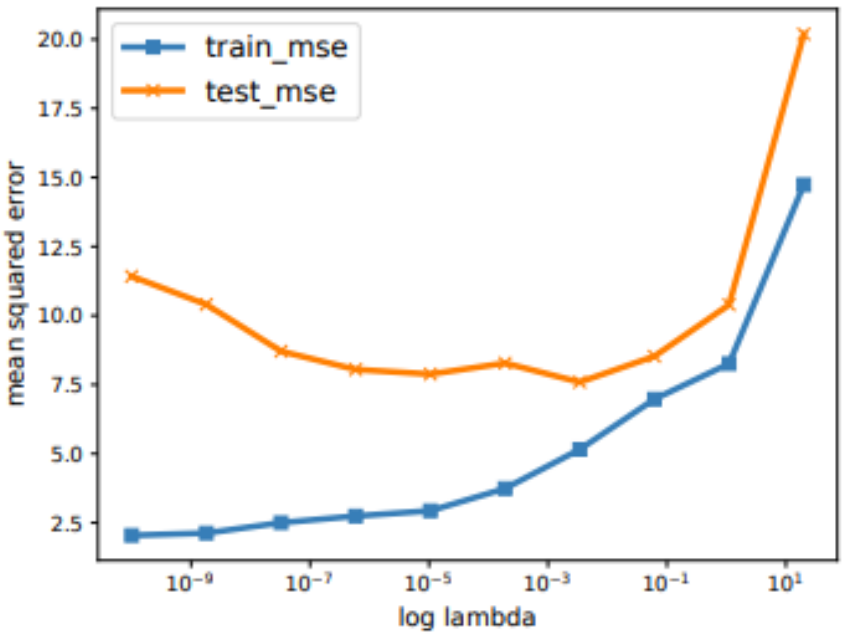
- Regularized empirical risk:

$$R_\lambda(\boldsymbol{\theta}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \ell(\boldsymbol{y}, f(\boldsymbol{x};\boldsymbol{\theta})) + \lambda C(\boldsymbol{\theta})$$

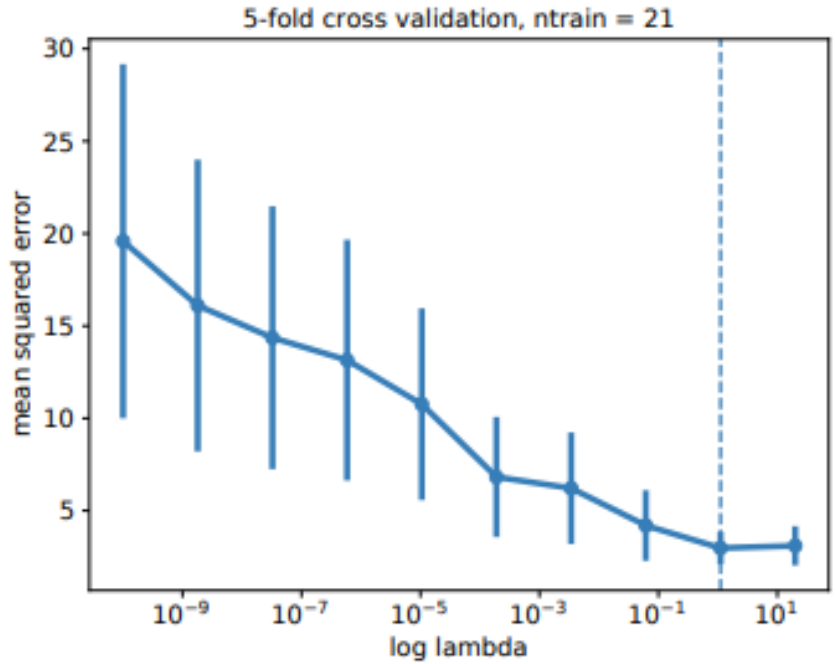
- We split the training data into K folds; then, for each fold $k \in \{1, \ldots, K\}$, we train on all the folds but the k'th:

$$R_\lambda^{\mathrm{cv}} \triangleq \frac{1}{K} \sum_{k=1}^{K} R_0 \left(\hat{\boldsymbol{\theta}}_\lambda \left(\mathcal{D}_{-k} \right), \mathcal{D}_k \right)$$

- $\hat{\lambda} = \mathrm{argmin}_\lambda \, R_\lambda^{\mathrm{cv}}$.
- $\hat{\boldsymbol{\theta}} = \mathrm{argmin}_{\boldsymbol{\theta}} \, R_{\hat{\lambda}}(\boldsymbol{\theta}, \mathcal{D})$.
- We can use the **one-standard error rule** to select a simpler model.


train_mse
test_mse
mean squared error
log lambda
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
10^{-9}
10^{-7}
10^{-5}
10^{-3}
10^{-1}
10^{1}

(a)


5-fold cross validation, ntrain = 21
mean squared error
log lambda
30
25
20
15
10
5
10^{-9}
10^{-7}
10^{-5}
10^{-3}
10^{-1}
10^{1}

(b)

Table of Contents

Bayesian Statistics

- In statistics, modeling uncertainty about parameters using a probability distribution is known as inference.
- Bayesian statistics: using **posterior distribution** to represent uncertainty about parameters.
- Computing posterior distribution (Bayes rule):

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\boldsymbol{\theta})p(\mathcal{D} \mid \boldsymbol{\theta})}{p(\mathcal{D})} = \frac{p(\boldsymbol{\theta})p(\mathcal{D} \mid \boldsymbol{\theta})}{\int p\left(\boldsymbol{\theta}'\right) p\left(\mathcal{D} \mid \boldsymbol{\theta}'\right) d\boldsymbol{\theta}'}$$

  - conjugate prior + exponential family $\rightarrow$ closed-form posterior.
  - general cases $\rightarrow$ numerical methods (Laplace approximation, VI, MCMC, ...).
- Posterior predictive distribution (Bayes model averaging):

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \mathcal{D}) = \int p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}) d\boldsymbol{\theta}$$

Example: Bernoulli model

- We toss a coin N times, and want to infer the probability of heads.
- $\mathcal{D} = \{y_n : n = 1 : N\}$ is all the data. We assume $y_n \sim \text{Ber}(\theta)$.
- Prior: $p(\theta) = \text{Beta}(\theta \mid \breve{\alpha}, \breve{\beta}) = \frac{1}{B(\breve{\alpha},\breve{\beta})} \theta^{\breve{\alpha}-1}(1-\theta)^{\breve{\beta}-1}$.
- Likelihood: $p(\mathcal{D} \mid \theta) = \prod_{n=1}^{N} \text{Ber}(y_n \mid \theta) = \theta^{N_1}(1-\theta)^{N_0}$.
- Posterior:

$$\begin{aligned} p(\theta \mid \mathcal{D}) &\propto \theta^{N_1}(1-\theta)^{N_0}\theta^{\breve{\alpha}-1}(1-\theta)^{\breve{\beta}-1} \\ &\propto \text{Beta}\left(\theta \mid \breve{\alpha} + N_1, \breve{\beta} + N_0\right) \\ &= \text{Beta}(\theta \mid \hat{\alpha}, \hat{\beta}) \end{aligned}$$

- If we set $\breve{\alpha} = \breve{\beta} = 1$, then $p(\theta) = \text{Unif}(\theta \mid 0, 1)$. (uninformative prior)

Example: Bernoulli model (posterior distribution)

- Posterior Mode:

$$\begin{aligned}\hat{\theta}_{\text{map}} &= \arg\max_{\theta} p(\theta \mid \mathcal{D}) \\ &= \arg\max_{\theta} \log p(\theta \mid \mathcal{D}) \\ &= \arg\max_{\theta} \log p(\theta) + \log p(\mathcal{D} \mid \theta)\end{aligned}$$

- Notice that if we use a uniform prior, then the MAP estimate is the same as the MLE estimate. (another justification of MLE)
- $p(\theta) = \text{Beta}(\theta \mid \breve{\alpha}, \breve{\beta})$:

$$\hat{\theta}_{\text{map}} = \frac{\breve{\alpha} + N_1 - 1}{\breve{\alpha} + N_1 - 1 + \breve{\beta} + N_0 - 1}$$

- If $p(\theta) = \text{Beta}(\theta|2, 2)$, then $\hat{\theta}_{\text{map}} = \frac{N_1+1}{N+2}$. (add-one smoothing)
- If $p(\theta) = \text{Beta}(\theta|1, 1)$, then $\hat{\theta}_{\text{map}} = \frac{N_1}{N}$. (MLE)

- Posterior Mean: $\bar{\theta} \triangleq \mathbb{E}[\theta \mid \mathcal{D}] = \frac{\hat{\alpha}}{\hat{\beta}+\hat{\alpha}} = \frac{\hat{\alpha}}{\hat{N}}$.
  - Define prior strength $\breve{N} \triangleq \breve{\alpha} + \breve{\beta}$.
  - Ratio of the prior to posterior equivalent sample size: $\lambda = \breve{N}/N$.
  - Prior mean: $m = \breve{\alpha}/\breve{N}$.
  - Posterior mean: $\bar{\theta} = \lambda m + (1-\lambda)\hat{\theta}_{mle}$.
- Posterior Variance: $\mathrm{Var}[\theta \mid \mathcal{D}] = \frac{\hat{\alpha}\hat{\beta}}{(\hat{\alpha}+\hat{\beta})^2(\hat{\alpha}+\hat{\beta}+1)}$.
  - If $N \gg \breve{N}$, $\sigma = \sqrt{\mathbb{V}[\theta \mid \mathcal{D}]} \approx \sqrt{\frac{\hat{\theta}_{mle}(1-\hat{\theta}_{mle})}{N}}$
  - Uncertainty goes down at a rate of $1/\sqrt{N}$.
  - The uncertainty is maximized when $\hat{\theta}_{mle} = 0.5$, i.e., it's easier to be sure a coin is biased than fair.

Example: Bernoulli model (posterior predictive)

- Assume we use the uniform prior.
- We flip the coin three times, and we see three heads. Thus, $\hat{\theta}_{map} = \hat{\theta}_{mle} = 1$.
- What's the predicted probability that the next flip is head?
- Plug-in approximation:
  - $p(y = 1|\hat{\theta}) = \text{Ber}(y = 1|\theta = 1) = 1$.
  - black swan problem!
- Posterior predictive distribution:

$$\begin{aligned} p(y = 1 \mid \mathcal{D}) &= \int_0^1 p(y = 1 \mid \theta) p(\theta \mid \mathcal{D}) d\theta \\ &= \int_0^1 \theta \, \text{Beta}(\theta \mid \hat{\alpha}, \hat{\beta}) d\theta \\ &= \mathbb{E}[\theta \mid \mathcal{D}] = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}} = \frac{1+3}{(1+3)+1} = 0.8. \end{aligned}$$

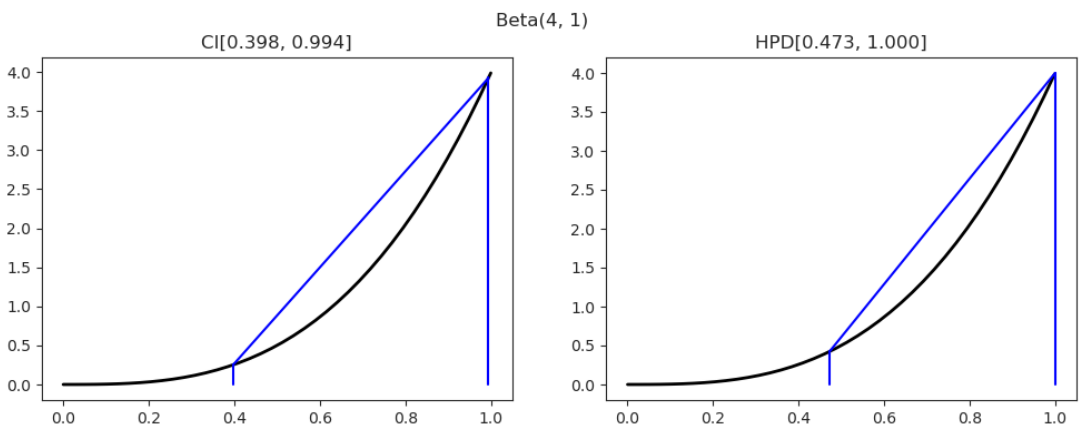
- We still have add-one smoothing even if we use uninformative prior.

Credible Interval

- We define a $100(1-\alpha)\%$ credible interval to be a (contiguous) region $C = (\ell, u)$ (standing for lower and upper) which contains $1-\alpha$ of the **posterior probability mass**, i.e.,

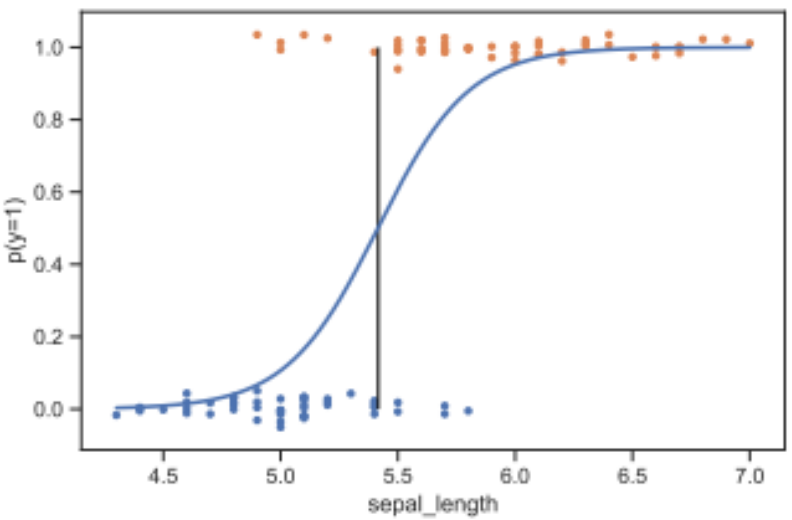
$$C_\alpha(\mathcal{D}) = (\ell, u) : P(\ell \leq \theta \leq u \mid \mathcal{D}) = 1 - \alpha$$

- central interval (CI):$\ell = F^{-1}(\alpha/2), u = F^{-1}(1-\alpha/2)$.
- highest posterior density (HPD)
  - A problem with central intervals is that there might be points outside the central interval that have a higher probability than points that are inside.
  - Find threshold p^* s.t. $1-\alpha = \int_{\theta: p(\theta|\mathcal{D}) > p^*} p(\theta \mid \mathcal{D}) d\theta$.
  - $C_\alpha(\mathcal{D}) = \{\theta : p(\theta \mid \mathcal{D}) \geq p^*\}$.



Bayesian machine learning

- Suppose we want to use logistic regression to classify Setosa and Versicolor.
- $p(y \mid \boldsymbol{x}; \boldsymbol{\theta}) = \text{Ber}\left(y \mid \boldsymbol{\sigma}\left(\boldsymbol{w}^\top \boldsymbol{x} + b\right)\right)$, where x is sepal length and $\boldsymbol{\sigma}$ is the sigmoid function.
- posterior distribution of $w, b \rightarrow$ credible interval of the decision boundary.



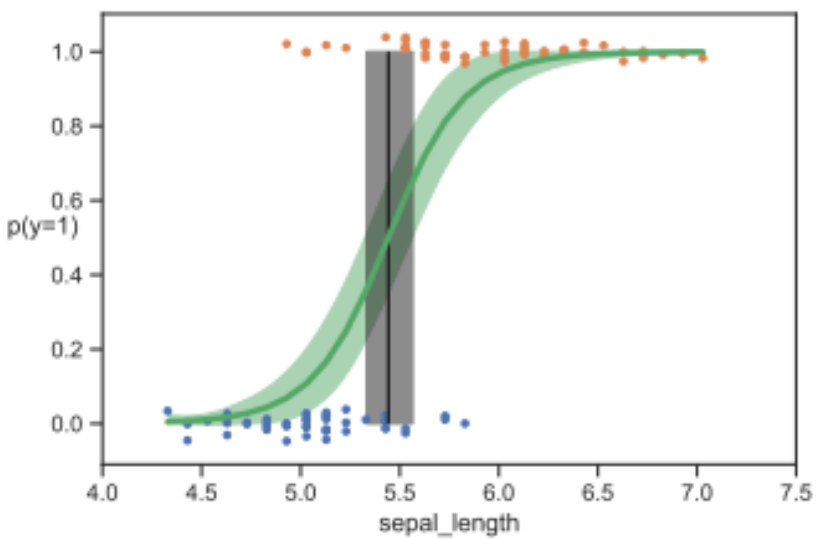


Table of Contents

Frequentist statistics v.s. Bayesian statistics

- Bayesian statistics: parameters are random variables.
- Frequentist statistics: parameters are fixed but unknown, and data is random. (no priors, no Bayes rule)
- It is this notion of variation across repeated trials that forms the basis for modeling uncertainty used by the frequentist approach.
- The Bayesian approach views probability in terms of information rather than repeated trials. (no asymptotic theory)
- *I believe that it would be very difficult to persuade an intelligent person that current [frequentist] statistical practice was sensible, but that there would be much less difficulty with an approach via likelihood and Bayes' theorem.* — George Box, 1962

Sampling distributions

- Estimator: $\hat{\theta} = \pi(\mathcal{D})$.
- Imagine sampling S different datasets, from true model $p(\boldsymbol{x}|\theta^*)$:

$$\mathcal{D}^{(s)} = \{\boldsymbol{x}_n \sim p\left(\boldsymbol{x}_n \mid \boldsymbol{\theta}^*\right) : n = 1 : N\}$$

- Let $S \to \infty$, we get the sampling distribution of the estimator $\hat{\theta}$:

$$p(\pi(\tilde{\mathcal{D}}) = \hat{\theta} \mid \tilde{\mathcal{D}} \sim \theta^*) = \lim_{S \to \infty} \frac{1}{S} \sum_{s=1}^{S} \delta\left(\hat{\theta} - \pi\left(\mathcal{D}^{(s)}\right)\right)$$

- Gaussian approximation of sampling distribution of MLE:

$$p(\pi(\tilde{\mathcal{D}}) = \hat{\theta} \mid \tilde{\mathcal{D}} \sim \theta^*) \to \mathcal{N}\left(\hat{\theta} \mid \boldsymbol{\theta}^*, \left(N\mathbf{F}\left(\boldsymbol{\theta}^*\right)\right)^{-1}\right)$$

  where F is the Fisher information matrix.
- Bootstrap: "poor man's posterior."

Confidence intervals

- We define a $100(1-\alpha)\%$ confidence interval for parameter θ as any interval $I(\tilde{\mathcal{D}}) = (\ell(\tilde{D}), u(\tilde{D}))$ derived from any $\tilde{D}$ such that:

$$\Pr\left(\theta \in I(\tilde{\mathcal{D}}) \mid \tilde{\mathcal{D}} \sim \theta\right) \geq 1-\alpha$$

- $\tilde{\mathcal{D}}, I(\tilde{\mathcal{D}})$ are random, θ is fixed!
- If we repeatedly sampled data, and compute $I(\tilde{D})$ for each such dataset, then about 95% of such intervals will contain the true parameter θ.
- It doesn't mean θ has 95% chance to live in a particular $I(\tilde{D})$.
- This counter-intuitive definition of confidence intervals can lead to bizarre results.

- Suppose we draw two integers $\mathcal{D} = (y_1, y_2)$ from

$$p(y \mid \theta) = \begin{cases} 0.5 & \text{if } y = \theta \\ 0.5 & \text{if } y = \theta + 1 \\ 0 & \text{otherwise} \end{cases}$$

  If $\theta = 39$, we would expect the following outcomes each with probability 0.25 : $(39, 39), (39, 40), (40, 39), (40, 40)$.
- Let $m = \min(y_1, y_2)$ and define the following interval:

$$[\ell(\mathcal{D}), u(\mathcal{D})] = [m, m]$$

  For the above samples this yields $[39, 39], [39, 39], [39, 39], [40, 40]$.
- Hence $[\ell(\mathcal{D}), u(\mathcal{D})]$ is a 75% CI.
- If we observe $\mathcal{D} = (39, 40)$, we know $\theta = 39$ for sure, yet we only have 75% "confidence" in this fact.

The bias-variance tradeoff

- mean of estimates: $\bar{\theta} = \mathbb{E}[\hat{\theta}]$.
- variance of estimates: $\mathrm{Var}[\hat{\theta}] \triangleq \mathbb{E}[\hat{\theta}^2] - (\mathbb{E}[\hat{\theta}])^2$.

$$
\begin{aligned}
\mathbb{E}\left[\left(\hat{\theta} - \theta^*\right)^2\right] &= \mathbb{E}\left[\left[(\hat{\theta} - \bar{\theta}) + \left(\bar{\theta} - \theta^*\right)\right]^2\right] \\
&= \mathbb{E}\left[(\hat{\theta} - \bar{\theta})^2\right] + 2\left(\bar{\theta} - \theta^*\right)\mathbb{E}[\hat{\theta} - \bar{\theta}] + \left(\bar{\theta} - \theta^*\right)^2 \\
&= \mathbb{E}\left[(\hat{\theta} - \bar{\theta})^2\right] + \left(\bar{\theta} - \theta^*\right)^2 \\
&= \mathbb{V}[\hat{\theta}] + \mathrm{bias}^2(\hat{\theta})
\end{aligned}
$$

- $\mathrm{MSE} = \text{variance} + \mathrm{bias}^2$.
- In the frequentist framework, MAP estimate uses a small bias to achieve a large decrease in variance, thus lower MSE compared to MLE.