Bayesian mutlimodeling: Bayesian inference

MIPT

2022

Coin problem

A man flips a coin N times. What's the probability of getting tails on a coin?

Coin problem

A man flips a coin 3 times. All 3 times it comes up tails. What's the probability of getting tails on a coin?

Naive approach

$$\boldsymbol{X} = [1, 1, 1];$$

$$x \sim \mathrm{Bin}(w);$$

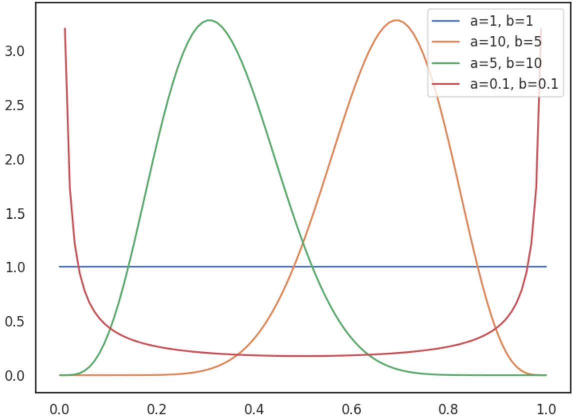
$$\hat{w} = \arg\max_{p} L(\boldsymbol{X}, w);$$

$$\rightarrow \hat{w} = 1.$$

Challenge: three events are not be enough to estimate the distribution of heads and tails.

Beta-distribution: recap

- corresponds to the *prior* beliefs about Bernoulli distribution
- interpretation: "effective number of events $w = 1, w = 0$"
- With $n \to \infty$ converges to δ-distribtuion with PDF concentration at MLE for Bernoulli distribution.



Bayesian approach

Use beta-distribution as a *prior* distribution for our parameter w. From general considerations, the distribution should be symmetrical (unless we have more information):

$$p(w) \sim B(\alpha, \beta).$$

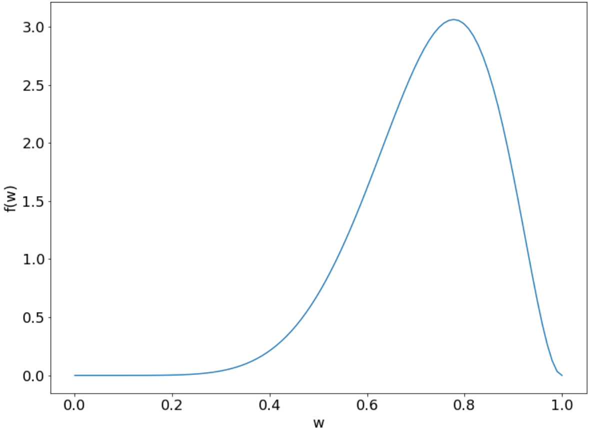
Find the *posterior* distribution of w using Bayes formula:

$$p(w|\boldsymbol{x}) = \frac{p(\boldsymbol{X}|w)p(w)}{p(\boldsymbol{X})} \propto p(\boldsymbol{X}|w)p(w);$$

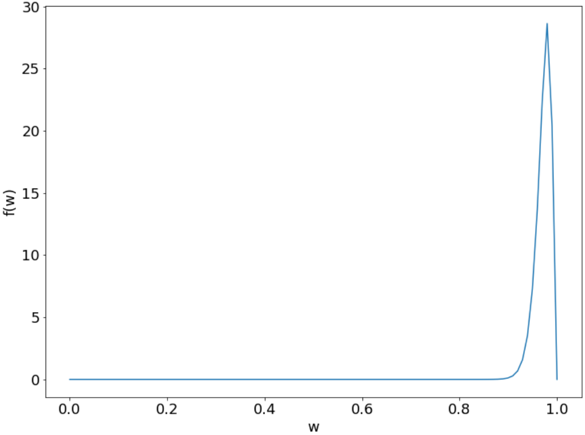
$$\log p(w|\boldsymbol{x}) = \log p(\boldsymbol{X}|w) + \log p(w) + \text{Const}.$$

Conclusion: roughly prior is a — regularizer.

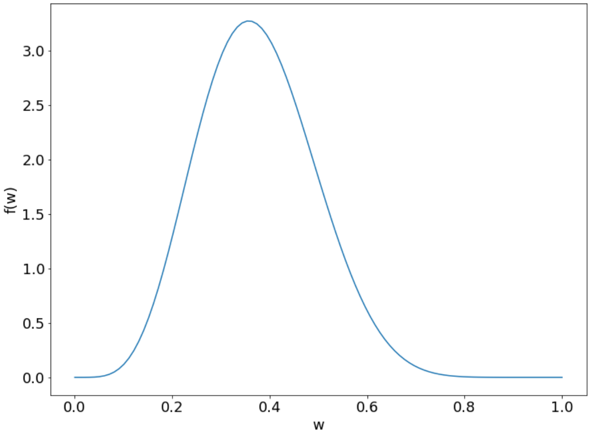
Posterior, $\alpha = 3, \beta = 3$



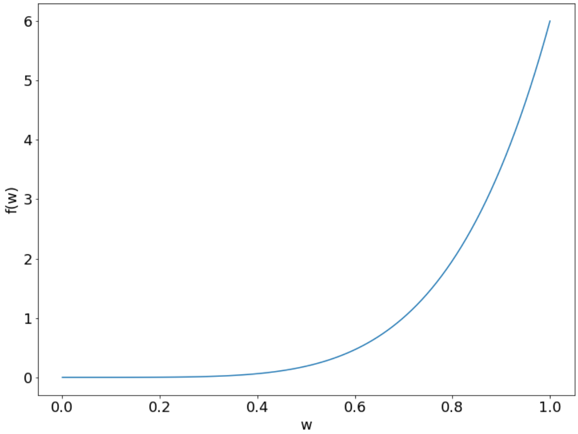
Posterior, 100 elements



Posterior, $\alpha = 1, \beta = 10$



Posterior, $\alpha = 1, \beta = 1$



Bayesian inference: first level

Given:

- Likelihood $p(\boldsymbol{X}|\boldsymbol{w})$ of the dataset $\boldsymbol{X}$ w.r.t. parameters $\boldsymbol{w}$;
- prior distribution $p(\boldsymbol{w}|\boldsymbol{h})$
- prior parameters $\boldsymbol{h}$ (for the coin problem: $\boldsymbol{h} = [\alpha, \beta]$;)

Then the posterior for $\boldsymbol{w}$ w.r.t. $\boldsymbol{X}$:

$$p(\boldsymbol{w}|\boldsymbol{x}, \boldsymbol{h}) = \frac{p(\boldsymbol{X}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{h})}{p(\boldsymbol{X}|\boldsymbol{h})} \propto p(\boldsymbol{X}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{h}).$$

Find a point estimate as a maximum posterior probability (MAP):

$$\hat{\boldsymbol{w}} = \arg\max p(\boldsymbol{X}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{h}).$$

MAP-estimation is similar to MLE, if

- the dataset is large;
- prior is uniform in an infinitely large region (improper prior)

Why we used Beta-distribution?

$$p(w|\boldsymbol{x}, \alpha, \beta) \propto p(\boldsymbol{X}|w)p(w|\alpha, \beta) \propto$$
$$\propto w^{\sum x}(1-w)^{m-\sum x} \times w^{\alpha-1}(1-w)^{\beta-1} =$$
$$= w^{\alpha-1+\sum x}(1-w)^{m+\beta-\sum x-1} \sim B(\alpha+\sum x, \beta+m-\sum x).$$

The distribution family is conjugate prior to the likelihood distribution, if the posterior belongs to the same family.

Prior families

- Discrete (labels, discrete parameters)
  - Bernoulli
  - Categorial distributions

  Hyperparameters (parameters of the prior parameters):
  - $w \sim \text{Bin}(w)$: $w \sim B(\alpha, \beta)$: conjugate
  - $w \sim \text{Cat}(w)$: $w \sim \text{Dir}(\boldsymbol{\alpha})$: conjugate
- Real-valued distributions
  - $\mathcal{N}$
  - Laplace
  - $\mathcal{C}$

  Hyperparameters:
  - Variance, $w \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2), \sigma^{-1} \in \Gamma$: conjugate for Gaussian distribution
  - Expectation, $\boldsymbol{\mu} \in \mathcal{N}$: conjugate for Gaussian distribution

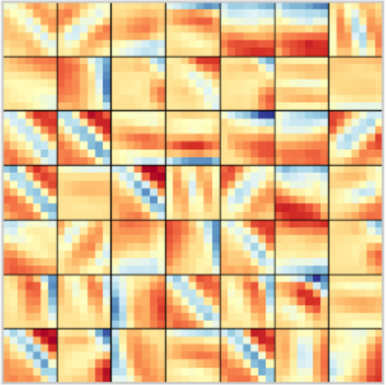
Informative prior vs Uninformative prior

- Informative prior: corresponds to some expert knowledge
  - Example: air temperature in some region: Gaussian variable with known mean and variance estimated from previous observations.
  - Mistanke in informative prior estimation leads to poor models.
- Uninformative prior: corresponds to some basic knowledge
  - Example: air temperature in some region: uniform improper prior.
- Weakly-informative prior: somewhere in between
  - Example: air temperature in some region: uniform distribution in [-50, 50].

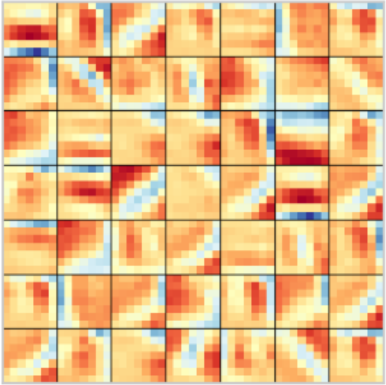
To discuss:

- $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{A}^{-1})$ — what type of the prior distribution?
- What if our prior and posterior are very close?

The deep weight prior: Atanov et al., 2019



(b) Learned filters



(c) Samples from DWP

The distribution can be modeled by complex models and can generate rather informative samples!

Jeffreys prior

Uninformative prior:

$$p(\mathbf{w}) \propto \sqrt{\det I(\mathbf{w})} = \sqrt{\det\left(-\frac{\partial^2}{\partial w^2}\log L(w)\right)}.$$

- Invariant under to the variable change:

$$p(g(\mathbf{w})) = p(\mathbf{w})\left|\frac{dg}{d\mathbf{w}}\right| \to$$

$$p(g(\mathbf{w})) \propto \sqrt{\det I(g(\mathbf{w}))}.$$

- Interpretation: a value inverse to the amount of information obtained by our model from the dataset
- Examples:
  - $y \in \mathrm{Bin}(w)$: $\quad p(w) \propto \frac{1}{\sqrt{p(1-p)}}$ — Beta-distribution (0.5, 0.5).
  - $w \in \mathcal{N}(\mu, \sigma)$: $\quad p(\mu) \propto \mathrm{Const}.$
  - $w \in \mathcal{N}(\mu, \sigma)$: $\quad p(\sigma) \propto \frac{1}{|\sigma|}$.

Model selection problem: Bayesian coherent inference

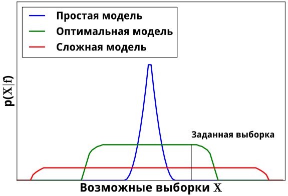
First level: find optimal parameters:

$$\mathbf{w} = \arg\max \frac{p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})}{p(\mathfrak{D}|\mathbf{h})},$$

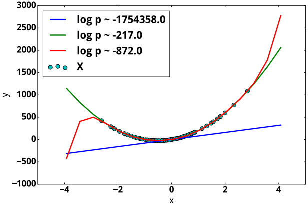
Second level: find model, that givew optimal Evidence.

Evidence:

$$p(\mathfrak{D}|\mathbf{h}) = \int_{\mathbf{w}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})d\mathbf{w}.$$



Model selection scheme



Example: polynomial regression

Example: linear regression

Given m objects with n features: $\boldsymbol{f}(\boldsymbol{X}, \boldsymbol{w}) = \boldsymbol{X}\boldsymbol{w}$; $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{f}(\boldsymbol{X}, \boldsymbol{w}), \beta^{-1})$, $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{A}^{-1})$.
Write down the integral:

$$p(\mathfrak{D}|\boldsymbol{h}) = p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{A}, \beta) = \frac{\sqrt{\beta \cdot |\boldsymbol{A}|}}{\sqrt{(2\pi)^{m+n}}} \int_{\boldsymbol{w}} \exp\left(-0.5\beta(\boldsymbol{y} - \boldsymbol{f})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{f})\right) \exp\left(-0.5\boldsymbol{w}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{w}\right) d\boldsymbol{w} =$$

$$= \frac{\sqrt{\beta \cdot |\boldsymbol{A}|}}{\sqrt{(2\pi)^{m+n}}} \int_{\boldsymbol{w}} \exp(-S(\boldsymbol{w}))d\boldsymbol{w}$$

Its value is tractable for the linear regression case:

$$\int_{\boldsymbol{w}} \exp(-S(\boldsymbol{w}))d\boldsymbol{w} = (2\pi)^{\frac{n}{2}} \exp(-S(\hat{\boldsymbol{w}}))|\boldsymbol{H}^{-1}|^{0.5},$$

where

$$\boldsymbol{H} = \boldsymbol{A} + \beta\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X},$$

$$\hat{\boldsymbol{w}} = \beta\boldsymbol{H}^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}$$

Conclusion: we can find the value of the Evidence for the linear models.

Example: Laplace approximation

Given m objects with n features: $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{f}(\boldsymbol{X}, \boldsymbol{w}), \beta^{-1}), \boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{A}^{-1})$.
Write down the integral:

$$p(\mathfrak{D}|\boldsymbol{h}) = p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{A}, \beta) = \frac{\sqrt{\beta \cdot |\boldsymbol{A}|}}{\sqrt{(2\pi)^{m+n}}} \int_{\boldsymbol{w}} \exp(-S(\boldsymbol{w}))d\boldsymbol{w}.$$

Use Taylor serioes for S:

$$S(\boldsymbol{w}) \approx S(\hat{\boldsymbol{w}}) + \frac{1}{2}\Delta\boldsymbol{w}^{\mathsf{T}}\boldsymbol{H}\Delta\boldsymbol{w}$$

Then:

$$\frac{\sqrt{\beta \cdot |\boldsymbol{A}|}}{\sqrt{(2\pi)^{m+n}}} S(\hat{\boldsymbol{w}}) \int_{\boldsymbol{w}} \exp(-\frac{1}{2}\Delta\boldsymbol{w}^{\mathsf{T}}\boldsymbol{H}\Delta\boldsymbol{w})d\boldsymbol{w}$$

The expression corresponds to the PDF for unnormalized Gaussian distribution.
Conclusion: we can use Laplace approximation for the non-linear models.

Laplace approximation: drawbacks

- Only Gaussian distribution is available
  - No multimodality
- Hessian inversion: terribely slow
  - we can use diagonal matrix, but with worse approximation

A scalable Laplace approximation for neural networks: Ritter et al., 2018

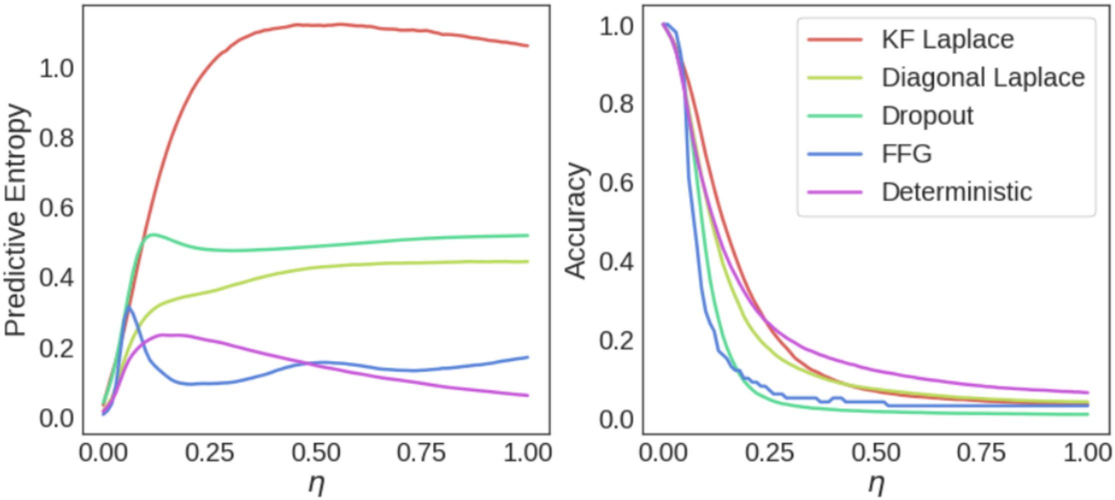
- Decompose the neural network parameters by the layers, make an assumption that parameters from different layers are not correlated
- $\boldsymbol{H}_l = (\boldsymbol{a}_i \boldsymbol{a}_i^\top) \circ \boldsymbol{H}(\boldsymbol{h}_l)$, with Kronecker product, where for layer l:

$$\boldsymbol{a}_l = \sigma(\mathrm{W}_l \boldsymbol{a}_{l-1}) = \sigma(\boldsymbol{h}_l)$$

.

- Reduce the complexity from d^4 to d^2 (if the dimmension is d)
- Kronecker product is equal to the Kronecker product of the inverses

Approximation mode matters



References

- Bishop C. M. Pattern recognition //Machine learning. – 2006. – T. 128. – №. 9.
- MacKay D. J. C., Mac Kay D. J. C. Information theory, inference and learning algorithms. – Cambridge university press, 2003.
- Atanov A. et al. The deep weight prior //7th International Conference on Learning Representations, ICLR 2019. – 2019.
- Kuznetsov M., Tokmakova A., Strijov V. Analytic and stochastic methods of structure parameter estimation //Informatica. – 2016. – T. 27. – №. 3. – C. 607-624.
- Coin problem: https://towardsdatascience.com/visualizing-beta-distribution-7391c18031f1
- Jeffreys prior: https://medium.datadriveninvestor.com/firths-logistic-regression-classification-with-datasets-that-are-small-imbalanced-or-separated-49d7782a13f1
- Ritter H., Botev A., Barber D. A scalable laplace approximation for neural networks //6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings. – International Conference on Representation Learning, 2018. – T. 6.
- Aleksandr Kolesov talk: https://github.com/intsystems/BMM/blob/main-21/