ISTA 421/521 Introduction to Machine Learning

Lecture 12: Marginal Likelihood and Hyperparameters

Clay Morrison
clayton@sista.arizona.edu
Gould-Simpson 819
Phone 621-6609

2 October 2014

Which Prior is the Correct One?

- Well, it depends. Sometimes it is justified by background knowledge and context.
- As we get new data, the effect of the prior diminishes.
- Another approach: Look at the marginal likelihoods

Marginal Likelihood

- $P(y_N)$ is the marginal probability of the data. It can be related to r:

$$p(y_N) = \int_{r=0}^{r=1} p(r, y_N)\, dr \;=\; \int_{r=0}^{r=1} p(y_N|r)p(r)\, dr$$

- Need to be explicit about conditioning of r:

$$p(y_N|\alpha,\beta) = \int_{r=0}^{r=1} p(y_N|r)p(r|\alpha,\beta)\, dr$$

- This tells us how likely the data is given our choice of prior parameters α and β.
- The higher $p(y_N|\alpha,\beta)$, the better our evidence agrees with the prior specification.
- Can use $p(y_N|\alpha,\beta)$ to help choose best scenario: choose scenario with highest $p(y_N|\alpha,\beta)$.

SISTA

Evaluate the Marginal Likelihood Integral

$$\begin{aligned} p(y_N|\alpha,\beta) &= \int_{r=0}^{r=1} p(y_N|r)p(r|\alpha,\beta)\, dr \\ &= \int_{r=0}^{r=1} \left(\begin{array}{c} N \\ y_N \end{array}\right) r^{y_N}(1-r)^{N-y_N} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}\, dr \\ &= \left(\begin{array}{c} N \\ y_N \end{array}\right) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_{r=0}^{r=1} r^{\alpha+y_N-1}(1-r)^{\beta+N-y_N-1}\, dr. \end{aligned}$$

We've dealt with this integration problem before:

$$\int_{r=0}^{r=1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1} dr = 1 \qquad \int_{r=0}^{r=1} r^{\alpha-1}(1-r)^{\beta-1} dr = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$p(y_N|\alpha,\beta) = \left(\begin{array}{c} N \\ y_N \end{array}\right) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_N)\Gamma(\beta+N-y_N)}{\Gamma(\alpha+\beta+N)}$$

SISTA

Evaluate the Marginal Likelihood for the different scenarios

$$p(y_N|\alpha,\beta) = \binom{N}{y_N}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+y_N)\Gamma(\beta+N-y_N)}{\Gamma(\alpha+\beta+N)}$$

In our example, $N = 20$ and $y_N = 14$

Observations:

H T H H H H T T T H

H H T T H H H H H H

1. No prior knowledge, $\alpha = \beta = 1$, $p(y_N \mid \alpha, \beta) = 0.0476$
2. Fair coin, $\alpha = \beta = 50$, $p(y_N \mid \alpha, \beta) = 0.0441$ ← lowest
3. Biased coin, $\alpha = 5$, $\beta = 1$, $p(y_N \mid \alpha, \beta) = 0.0576$ ← highest

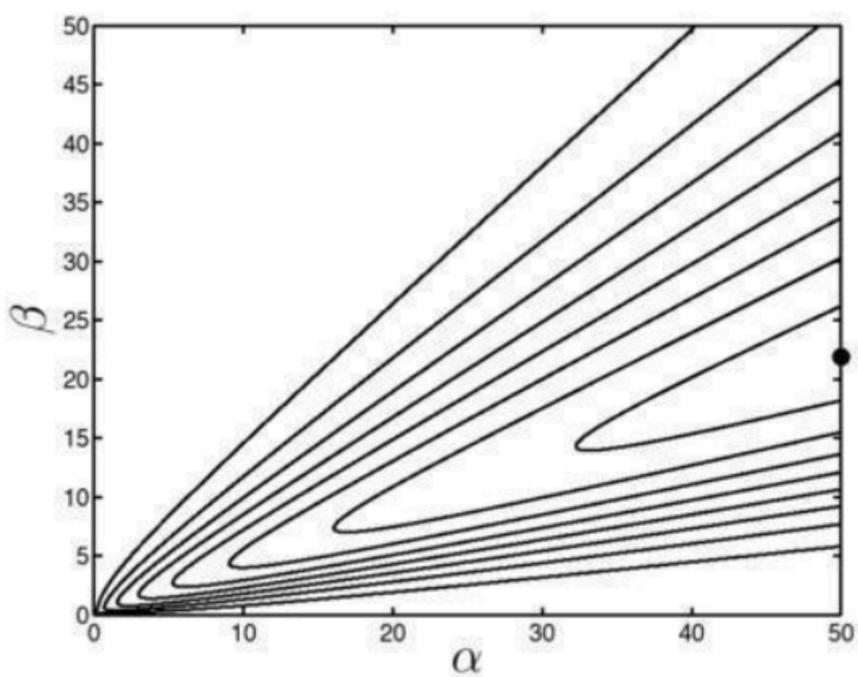
Caution: Choosing this way makes the prior **no longer** correspond to our beliefs ***before*** we observe any data.

AKA: Type II Maximum Likelihood

Optimize α and β using Marginal Likelihood

$0 \leq \alpha \leq 50$

$0 \leq \beta \leq 30$



Treating Parameters as R.V.s

- In some cases we have good reason to select particular parameter values based on knowledge.
- Other times, we don't know the exact value, so treat as random variables themselves.
- Often useful and appropriate to treat as independent, and capitalize on conditional independence
- E.g., prior density over all random variables

$$p(r, \alpha, \beta) = p(r|\alpha, \beta)p(\alpha, \beta)$$

- For our model, we want the posterior over all parameters in our model

$$\begin{aligned} p(r, \alpha, \beta | y_N) &= \frac{p(y_N|r, \alpha, \beta)p(r, \alpha, \beta)}{p(y_N)} \\ &= \frac{p(y_N|r)p(r, \alpha, \beta)}{p(y_N)} \quad \text{Conditional independence} \\ &= \frac{p(y_N|r)p(r|\alpha, \beta)p(\alpha, \beta)}{p(y_N)} \end{aligned}$$

SISTA

R.V. Parameters may have…

Hyperparameters!

- *κ* controls the density in the same way that *α* and β control the density for *r*

$$p(\alpha, \beta | \kappa)$$

- When computing marginal likelihood: integrate over all random variables, leaves us with the data conditioned on the hyper-parameters:

$$p(y_N|\kappa) = \int\int\int p(y_N|r)p(r|\alpha, \beta)p(\alpha, \beta|\kappa)\; dr\; d\alpha\; d\beta$$

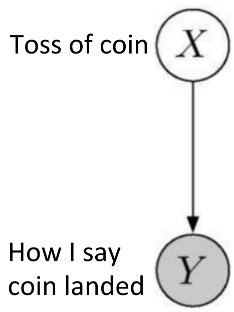
… can keep going: hierarchical models

SISTA

Graphical Models

… are a notation to compactly describe a complex relation of random variables:

nodes = RVs, edges = RV dependencies



$P(Y = y|\ X = x)$

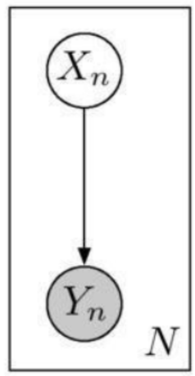
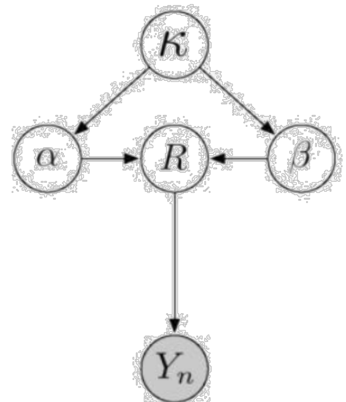


Plate Notation



Return (again) to the Olympics 100m

The Bayesian treatment…

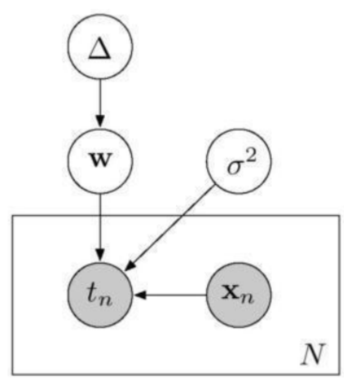
- First, the model:

$$t_n = w_0 + w_1 x_n + w_2 x_n^2 + \ldots + w_k x_n^k + \epsilon_n$$

k-th-order polynomial (Ch 1)
Gaussian distributed noise (Ch 2) $\quad \epsilon \sim \mathcal{N}(0, \sigma^2)$

$$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \qquad \mathbf{t} = \mathbf{X}^\top \mathbf{w} + \epsilon$$

$$\begin{aligned} p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \Delta) &= \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2, \Delta)p(\mathbf{w}|\Delta)}{p(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta)} \\ &= \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2)p(\mathbf{w}|\Delta)}{p(\mathbf{t}|\mathbf{X}, \sigma^2, \Delta)}. \\ &= \frac{p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2)p(\mathbf{w}|\Delta)}{\int p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma^2)p(\mathbf{w}|\Delta)\ d\mathbf{w}} \end{aligned}$$



Predictions, Likelihood & Prior

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\Delta) = \frac{p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)p(\mathbf{w}|\Delta)}{\int p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)p(\mathbf{w}|\Delta)\; d\mathbf{w}}$$

Can use $p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\Delta)$ to make predictions:

$$p(t_{new}|\mathbf{x}_{new},\mathbf{X},\mathbf{t},\sigma^2,\Delta) = \int p(t_{new}|\mathbf{x}_{new},\mathbf{w},\sigma^2)p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\Delta)\; d\mathbf{w}$$

$$p(t_{new} < 9.5|\mathbf{x}_{new},\mathbf{X},\mathbf{t},\sigma^2,\Delta) = \int p(t_{new} < 9.5|\mathbf{x}_{new},\mathbf{w},\sigma^2)p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\Delta)\, d\mathbf{w}$$

The Likelihood:

$p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2) = \mathcal{N}(\mathbf{X}\mathbf{w},\sigma^2\mathbf{I}_N)$ analogous to Binomial likelihood in coin example

The Prior:

Want an exact posterior, so want prior that is **conjugate** to the Gaussian likelihood

$p(\mathbf{w}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) = \mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)$ analogous to Beta prior in coin example

$$p(\mathbf{w}|\mathbf{t},\,\mathbf{X},\,\sigma^{\,2},\,\boldsymbol{\mu}_0,\,\boldsymbol{\Sigma}_0)$$

The Posterior

We know that a Gaussian prior for the mean (weights **w**) is conjugate with a Gaussian likelihood, so the posterior is Gaussian!
Our goal is therefore to multiply the two and manipulate the prior and likelihood to get them into a single Gaussian form.

Ignore any term that doesn't involve **w**

$$\begin{aligned}
p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2) &\propto p(\mathbf{t}|\mathbf{w},\mathbf{X},\sigma^2)p(\mathbf{w}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) \\
&= \frac{1}{(2\pi)^{N/2}|\sigma^2\mathbf{I}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{t}-\mathbf{X}\mathbf{w})^{\mathsf{T}}(\sigma^2\mathbf{I})^{-1}(\mathbf{t}-\mathbf{X}\mathbf{w})\right) \\
&\quad\times\frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}_0|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}(\mathbf{w}-\boldsymbol{\mu}_0)\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{t}-\mathbf{X}\mathbf{w})^{\mathsf{T}}(\mathbf{t}-\mathbf{X}\mathbf{w})\right)\exp\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}(\mathbf{w}-\boldsymbol{\mu}_0)\right) \\
&= \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}(\mathbf{t}-\mathbf{X}\mathbf{w})^{\mathsf{T}}(\mathbf{t}-\mathbf{X}\mathbf{w})+(\mathbf{w}-\boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}(\mathbf{w}-\boldsymbol{\mu}_0)\right)\right\}. \\
&\propto \exp\left\{-\frac{1}{2}\left(-\frac{2}{\sigma^2}\mathbf{t}^{\mathsf{T}}\mathbf{X}\mathbf{w}+\frac{1}{\sigma^2}\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{w}+\mathbf{w}^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}\mathbf{w}-2\boldsymbol{\mu}_0^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}\mathbf{w}\right)\right\}
\end{aligned}$$

Note: simplifying by ignoring any terms not including **w**

The Posterior

$$\propto \exp\left\{-\frac{1}{2}\left(-\frac{2}{\sigma^2}\mathbf{t}^\top\mathbf{X}\mathbf{w} + \frac{1}{\sigma^2}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} + \mathbf{w}^\top\boldsymbol{\Sigma}_0^{-1}\mathbf{w} - 2\boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\mathbf{w}\right)\right\}$$

The form we want…

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{w}, \boldsymbol{\Sigma}_\mathbf{w})$$

$$\propto \exp\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_\mathbf{w})^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}(\mathbf{w}-\boldsymbol{\mu}_\mathbf{w})\right)$$

$$\propto \exp\left\{-\frac{1}{2}\left(\underbrace{\mathbf{w}^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w}}_{\text{quadratic}} - \underbrace{2\boldsymbol{\mu}_\mathbf{w}^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w}}_{\text{linear}}\right)\right\}$$

Combine the quadratic terms…

$$\mathbf{w}^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w} = \frac{1}{\sigma^2}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} + \mathbf{w}^\top\boldsymbol{\Sigma}_0^{-1}\mathbf{w}$$

$$= \mathbf{w}^\top\left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)\mathbf{w}$$

$$\boxed{\boldsymbol{\Sigma}_\mathbf{w} = \left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}}$$

Combine the linear terms…

$$-2\boldsymbol{\mu}_\mathbf{w}^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w} = -\frac{2}{\sigma^2}\mathbf{t}^\top\mathbf{X}\mathbf{w} - 2\boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\mathbf{w}$$

$$\boldsymbol{\mu}_\mathbf{w}^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}\mathbf{w} = \frac{1}{\sigma^2}\mathbf{t}^\top\mathbf{X}\mathbf{w} + \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\mathbf{w}$$

$$\boldsymbol{\mu}_\mathbf{w}^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1} = \frac{1}{\sigma^2}\mathbf{t}^\top\mathbf{X} + \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}$$

$$\boldsymbol{\mu}_\mathbf{w}^\top\boldsymbol{\Sigma}_\mathbf{w}^{-1}\boldsymbol{\Sigma}_\mathbf{w} = \left(\frac{1}{\sigma^2}\mathbf{t}^\top\mathbf{X} + \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{\Sigma}_\mathbf{w}$$

$$\boldsymbol{\mu}_\mathbf{w}^\top = \left(\frac{1}{\sigma^2}\mathbf{t}^\top\mathbf{X} + \boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}_0^{-1}\right)\boldsymbol{\Sigma}_\mathbf{w}$$

$$\boldsymbol{\Sigma}_\mathbf{w}^\top = \boldsymbol{\Sigma}_\mathbf{w}$$

$$\boxed{\boldsymbol{\mu}_\mathbf{w} = \boldsymbol{\Sigma}_\mathbf{w}\left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)}$$

The Posterior

In summary:

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{w}, \boldsymbol{\Sigma}_\mathbf{w})$$

$$\boldsymbol{\Sigma}_\mathbf{w} = \left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}$$

$$\boldsymbol{\mu}_\mathbf{w} = \boldsymbol{\Sigma}_\mathbf{w}\left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{t} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right)$$

$$\boldsymbol{\mu}_\mathbf{w} = \left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{t} + \boldsymbol{\Sigma}_\mathbf{0}^{-1}\boldsymbol{\mu}_\mathbf{0}\right)$$

If $\boldsymbol{\mu}_0 = [0, 0, ..., 0]^\top$ then,

$$= \left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{t}$$

$$\boxed{\widehat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X} + N\lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{t}}$$

Given that the posterior is a Gaussian, the single most likely value of **w** is the mean of the posterior, $\boldsymbol{\mu}_\mathrm{w}$

This is the **maximum a posteriori** (MAP) estimate of **w**
… and is the maximum of (the product of the likelihood and the prior):

$$p(\mathbf{w}, \mathbf{t}|\mathbf{X}, \sigma^2, \Delta)$$

Now, an important connection: Recall that the squared loss considered in Chapter 1 is very similar to the Gaussian likelihood.
Computing the most likely posterior (when the likelihood is Gaussian) is equivalent to using **regularized least squares**!
This can help provide intuition about effect of prior:

inverse of prior covariance play a regularization role.