06-Linear Regression (Bayesian Learning)

Recap: MLE vs MAP parameter estimation

- **MLE:** Maximizing the likelihood of data. Unbiased if data are iid drawn from a linear model with zero-mean noise.
- **MAP**: Expectation over data is weird because data are known. MAP estimation defines the prior distribution of parameters and maximizes the posterior.
- Unfortunately, MAP estimation is deceptive Bayesian in that
  - The goal of machine learning is **NOT** to estimate parameters, but to **make a prediction**.
  - Therefore, we do not really care any particular parameters.

A Bayesian view

- Anything unknown is a random variable
- Anything unrelated to the ultimate task should be marginalized out

The general framework for Bayesian Learning

- We define the prior $p(\boldsymbol{\theta})$
- We compute the likelihood $p(\mathcal{D}|\boldsymbol{\theta})$
- The posterior is

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\boldsymbol{\theta})\,p(\mathcal{D}|\boldsymbol{\theta})}{p(\mathcal{D})} \propto_{\boldsymbol{\theta}} p(\boldsymbol{\theta})\cdot p(\mathcal{D}|\boldsymbol{\theta})$$

- Note again that we care more about prediction than parameters. The predictive distribution/density for a new data sample $\boldsymbol{x}_*$ is

$$\begin{aligned} p(t_*|\boldsymbol{x}_*, \mathcal{D}) &= \int p(t_*, \boldsymbol{\theta}|\boldsymbol{x}_*, \mathcal{D})\, d\boldsymbol{\theta} && \text{[marginalization]} \\ &= \int p(t_*|\boldsymbol{\theta}, x_*, \mathcal{D})\, p(\boldsymbol{\theta}|\boldsymbol{x}_*, \mathcal{D})\, d\boldsymbol{\theta} && \text{[conditional probability]} \\ &= \int p(t_*|\boldsymbol{\theta}, x_*)\, p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta} \end{aligned}$$

($\mathcal{D}$: Training data)

$[p(t_*|\boldsymbol{\theta}, x_*, \mathcal{D}) = p(t_*|\boldsymbol{\theta}, x_*)$ because $t_* \perp \mathcal{D}|\boldsymbol{\theta}$
$p(\boldsymbol{\theta}|x_*, \mathcal{D}) = p(\boldsymbol{\theta}|\mathcal{D})$ because $\boldsymbol{\theta} \perp x_*|\mathcal{D}]$

Bayesian Linear Regression

- We define the prior

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{m}_0, \mathbf{S}_0) \propto_{\mathbf{w}} \exp\left\{-\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^T \mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\right\}$$

- The likelihood of data, given parameters $\boldsymbol{w}$

$$p(\mathcal{D}|\mathbf{w}) \propto_{\mathbf{w}} \exp\left\{-\frac{1}{2}(\mathbf{t}-\mathbf{X}\mathbf{w})^T \boldsymbol{\Lambda}^{-1}(\mathbf{t}-\mathbf{X}\mathbf{w})\right\}$$

where $\boldsymbol{\Lambda} = \begin{pmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{pmatrix}$ if $t^{(m)} = \mathbf{w}^T \mathbf{x}^{(m)} + \varepsilon^{(m)}$ and $\varepsilon^{(m)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

- The posterior of $\boldsymbol{w}$ is

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathbf{w})\,p(\mathcal{D}|\mathbf{w})}{p(\mathcal{D})}$$

$$\propto_{\mathbf{w}} p(\mathbf{w})\,p(\mathcal{D}|\mathbf{w})$$

$$= \exp\left\{-\frac{1}{2}\left[\mathbf{w}^T(\mathbf{S}_0^{-1} + \mathbf{X}^T\boldsymbol{\Lambda}^{-1}\mathbf{X})\mathbf{w} - 2(\mathbf{m}_0^T\mathbf{S}_0^{-1} + \mathbf{y}^T\boldsymbol{\Lambda}^{-1}\mathbf{X})\mathbf{w} + \text{const}\right]\right\}$$

[Note, for symmetric matrix like $\mathbf{S}, \boldsymbol{\Lambda}$, $\mathbf{a}^T\mathbf{S}\mathbf{b} = \mathbf{b}^T\mathbf{S}\mathbf{a}$]

$$\propto_{\mathbf{w}} \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N)\right)$$

$$= \mathcal{N}(\mathbf{w}; \mathbf{m}_N, \mathbf{S}_N)$$

where $\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{t})$

$$\mathbf{S}_N = (\mathbf{S}_0^{-1} + \mathbf{X}^T\boldsymbol{\Lambda}^{-1}\mathbf{X})^{-1} = (\mathbf{S}_0^{-1} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X})^{-1}$$

- Comments
  - Easy-to-remember result:

    Gaussian · Gaussian = Gaussian

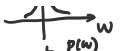
    Since the prior and the posterior take the same form (Gaussian distribution), it is known as a conjugate prior. Choosing a conjugate prior, unfortunately, is purely for computational convenience.
  - For an unbounded random variable like Gaussian, we cannot define uniform prior properly. Suppose we have no real prior information for the weights in linear regression, it's intuitive to assume

    $$\mathbf{m}_0 = \mathbf{0}, \quad \mathbf{S}_0 = \lambda^2\mathbf{I}$$

    In this case,

    $$\mathbf{m}_N = \frac{1}{\sigma^2}\mathbf{S}_N\mathbf{X}^T\mathbf{t}$$

    $$\mathbf{S}_N = (\frac{1}{\lambda^2}\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X})^{-1}$$

If, for the prior, $\lambda \to 0$, i.e.  , then $\mathbf{m}_N \to \mathbf{0} = \mathbf{m}_0$

If, for the prior, $\lambda \to \infty$, i.e.  , then $\mathbf{m}_N \to \hat{\mathbf{w}}_{MLE}$

No surprize, as was clear in MAP learning

If $\sigma \to 0$, i.e., data are less noisy, $\mathbf{m}_N \to \hat{\mathbf{w}}_{MLE}$

If $\sigma \to \infty$, i.e., data are very noisy, $\mathbf{m}_N \to \mathbf{0} = \mathbf{m}_0$

- The predictive density is

$$p(t_* \mid \boldsymbol{x}_*, \mathcal{D}) = \int p(t_* \mid \boldsymbol{x}_*, \mathbf{w})\, p(\mathbf{w} \mid \mathcal{D})\, d\mathbf{w}$$

$$= \int N(t_* \mid \mathbf{w}^T \boldsymbol{x}_*, \sigma^2) \cdot N(\mathbf{w} \mid \mathbf{m}_N, \mathbf{S}_N)\, d\mathbf{w}$$

$$\propto \int \exp\left\{-\frac{1}{2\sigma^2}(t_* - \mathbf{w}^T\boldsymbol{x}_*)^T(t_* - \mathbf{w}^T\boldsymbol{x}_*)\right\} \cdot \exp\left\{-\frac{1}{2}(\mathbf{w}-\mathbf{w}_N)^T \mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{w}_N)\right\} d\mathbf{w}$$

$$\simeq N(t_* \mid \mu_*, \sigma_*^2)$$

The observation of t_* following a normal distribution is due to the quadradic form in both $\mathbf{w}$ and y_*. By completing the square for $\mathbf{w}$, which integrates to 1, we still get a quadratic for t_*

where $\mu_* = \mathbf{m}_N^T \boldsymbol{x}_* = \boldsymbol{x}_*^T (\mathbf{X}^T\mathbf{X} + \frac{\lambda^2}{\sigma^2}\mathbf{I})^{-1}\mathbf{X}^T\mathbf{t}$

$\sigma_*^2 = \sigma^2 + \boldsymbol{x}_*^T \mathbf{S}_N \boldsymbol{x}_*$

- Bayesian linear regression
  - With zero-mean Gaussian prior, the mean prediction is the same as MAP
  - In addition, Bayesian linear regression also estimates a standard deviation, which reflects the model's degree of uncertainty.
  - This is useful for active learning, where the model could suggest data samples (in Bayesian linear regression, those with higher variance) for humans to label.

- An alternative way to compute the integral
  - **Monte Carlo (MC) sampling**

$$\mathop{\mathbb{E}}_{x \sim p(x)}[f(x)] = \int p(x) f(x)\, dx$$

$$\approx \frac{1}{K}\sum_{k=1}^{K} f(x^{(k)}) \quad \text{where} \quad x^{(k)} \overset{iid}{\sim} p(x)$$

The approximation error converges to zero if $M \to 0$ due to the Law of Large Numbers (LLN).

Note:

- The approximation holds for subjective probabilities because LLN is derived from probability axioms (which does not depend on how we interpret probabilities), and because intuitively, we can indeed sample infinitely many x's following our subjective probability, in which case, such sampling becomes a repeatable trial.
- The expectation can be thought of as weighted average, where the weights are probabilities. When we do MC sampling, the weights are implicit in how we obtain $x^{(1)}, \cdots, x^{(K)}$ and thus, MC sampling uses a direct average.
- The expectation can also be thought of as an "analytic" average, where we consider the precise probability of x taking a particular value. Thus, expectation alway gives an exact number (with "=" sign). MC sampling is an approximation with finite sampled values, and thus, always having the "≈" sign.

- Applying MC sampling to Bayesian learning

$$p(t_* \mid \boldsymbol{x}_*, \mathcal{D}) = \frac{1}{K} \sum_{k=1}^{K} p(t_* \mid \boldsymbol{x}_*, \boldsymbol{w}^{(k)})$$

where $\boldsymbol{w}^{(k)} \overset{iid}{\sim} p(\boldsymbol{w}^{(k)} \mid \mathcal{D})$

Choosing the prior

- For orthodox Bayesian, the prior is what a human believes. The prior is not subject to optimize.
  - Bayesian learning never overfits or underfits, because conditional Bayesian analysis, i.e., minimizing the Bayesian expected loss (over the posterior of parameters) is the only fundamentally correct decision principle. However, we may encounter the mis-fitting problem (similar to overfitting or underfitting) due to a bad prior.
- In practice, we may choose the prior by validation, or prior is learnable by
- Maximum likelihood estimation

Suppose $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \boldsymbol{\eta})$

$t \sim p(t \mid \boldsymbol{\theta})$ [conditioning on $\boldsymbol{x}$ is implicit]

$\boldsymbol{\eta}$ can be learned by MLE

$$\hat{\boldsymbol{\eta}} = \underset{\boldsymbol{\eta}}{\operatorname{argmax}}\ p(\boldsymbol{t} \mid \boldsymbol{\eta}) = \underset{\boldsymbol{\eta}}{\operatorname{argmax}} \int p(\boldsymbol{t} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{\eta})\, d\boldsymbol{\theta}$$

This is known as Type-II maximum likelihood (Type-II ML, or ML-II), or empirical Bayes.

- MAP-II

  Likewise, if we have a prior for $\boldsymbol{\eta}$, we could perform type-II max a posterior

$$\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} p(\boldsymbol{\eta} \mid \mathbf{t}) = \arg\max_{\boldsymbol{\eta}} \int p(\mathbf{t} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{\eta})\, p(\boldsymbol{\eta})\, d\boldsymbol{\theta}$$

- Full-Bayes (marginalization over $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$)

$$p(t_* \mid \boldsymbol{x}_*, \mathcal{D}) \propto \iint p(\boldsymbol{\eta}) \cdot p(\boldsymbol{\theta} \mid \boldsymbol{\eta}) \cdot p(\mathcal{D} \mid \boldsymbol{\theta}) \cdot p(t_* \mid \boldsymbol{x}_*, \boldsymbol{\theta})\, d\boldsymbol{\theta}\, d\boldsymbol{\eta}$$

- Final notes:
  - Theoretically, we may define a prior for the parameters of p($\boldsymbol{\eta}$), and this forms a deeper hierarchy of Bayes. In practice, we are happy with two-level prior.
  - Usually, we set the prior for $\boldsymbol{\eta}$ as the simplest or most intuitive one, e.g., standard norm or non-informative. Sometimes, a hierarchical prior may be more robust than one-level prior.
  - Technically, a hierarchical prior is merely a convenient vehicle for humans to express prior $p(\boldsymbol{\theta})$. It can be shown (how?) that a hierarchical prior can be collapsed to a usually prior, just like the usually parameter $\boldsymbol{\theta}$ collapses in ML-II and MAP-II.