Machine Learning and Computational Statistics
Bayesian Method and Regression

Intro: This document consists of concepts and exercises related to Bayesian Method and Regression. The core idea lies in the believe that distribution of parameters changes and we use Bayes' theorem to update the probability for a hypothesis (e.g., parameter distribution) as more evidence or information becomes available.

Learning objectives of Bayesian Method and Regression include:

1. Basic Bayesian setup (i.e., Posterior distribution = Prior distribution * Likelihood)
2. The prior predictive distribution:
$$p(y|x) = \int p(y|x;\theta)p(\theta)d\theta$$
3. The posterior predictive distribution:
$$p(y|x, D) = \int p(y|x;\theta)p(\theta|D)d\theta$$
4. The difference between the posterior predictive distribution function (Bayesian approach) and the MAP (Posterior Mean Estimate) (Frequentist approach)
   (a) Posterior predictive distribution does not depend on the unknown parameter because it has been integrated out.
   (b) Maximum a posteriori probability MAP: we choose $\hat{\theta}$ and predict $p(y|x, \hat{\theta}(D)$. We then take derivative of the log-likelihood and set to 0.
$$P(\theta|D) = \frac{P(D|\theta) * P(\theta)}{P(D)} \propto P(D|\theta) * P(\theta)$$
$$\begin{aligned}\hat{\theta}_{MAP} &= \underset{\theta}{\text{argmax}} P(D|\theta) * P(\theta) \\ &= \underset{\theta}{\text{argmax}} logP(D|\theta) * P(\theta) \\ &= \underset{\theta}{\text{argmax}} log \prod_{i=1}^{n} P(y_i|x_i, w) * P(\theta) \\ &= \underset{\theta}{\text{argmax}} \sum_{i=1}^{n} logP(y_i|x_i, w) * P(\theta)\end{aligned}$$

5. Conjugate prior:
   (a) Let π be a family of prior distributions on Θ
   (b) Let P be parametric family of distributions with parameter space Θ. P could be poisson, normal, or bernoulli
   (c) π is conjugate to P if for any prior in π, the posterior is always in π
6. Bayesian point estimates
   (a) posterior mean $\hat{\theta} = E(\theta|D)$
   (b) maximum a posteriori (MAP) estimator – the mode of the posterior distribution
7.

Bayesian Decision Theory

- Ingredients:
  - **Parameter space** Θ.
  - **Prior**: Distribution $p(\theta)$ on Θ.
  - **Action space** $\mathcal{A}$.
  - **Loss function**: $\ell : \mathcal{A} \times \Theta \to \mathbf{R}$.
- The **posterior risk** of an action $a \in \mathcal{A}$ is

$$\begin{aligned} r(a) &:= \mathbb{E}\left[\ell(\theta, a) \mid \mathcal{D}\right] \\ &= \int \ell(\theta, a) p(\theta \mid \mathcal{D})\, d\theta. \end{aligned}$$

  - It's the **expected loss under the posterior.**
- A **Bayes action** a^* is an action that minimizes posterior risk:

$$r(a^*) = \min_{a \in \mathcal{A}} r(a)$$

8. Relationship between Gaussian regression and Ridge regression

Closed Form for Posterior

- Model:

$$\begin{aligned} w &\sim \mathcal{N}(0, \Sigma_0) \\ y_i \mid x, w &\text{ i.i.d. } \mathcal{N}(w^T x_i, \sigma^2) \end{aligned}$$

- Design matrix X Response column vector y
- **Posterior distribution is a Gaussian distribution:**

$$\begin{aligned} w \mid \mathcal{D} &\sim \mathcal{N}(\mu_P, \Sigma_P) \\ \mu_{\mathbf{P}} &= \left(X^T X + \sigma^2 \Sigma_0^{-1}\right)^{-1} X^T y \\ \Sigma_{\mathbf{P}} &= \left(\sigma^{-2} X^T X + \Sigma_0^{-1}\right)^{-1} \end{aligned}$$

- **Posterior Variance** Σ_P gives us a natural **uncertainty measure.**

Closed Form for Posterior

- **Posterior distribution is a Gaussian distribution:**

$$
\begin{aligned}
w \mid \mathcal{D} &\sim \mathcal{N}(\mu_P, \Sigma_P) \\
\mu_{\mathbf{P}} &= \left(X^T X + \sigma^2 \Sigma_0^{-1}\right)^{-1} X^T y \\
\Sigma_{\mathbf{P}} &= \left(\sigma^{-2} X^T X + \Sigma_0^{-1}\right)^{-1}
\end{aligned}
$$

- If we want point estimates of w, **MAP estimator** and the **posterior mean** are given by

$$
\hat{w} = \mu_P = \left(X^T X + \sigma^2 \Sigma_0^{-1}\right)^{-1} X^T y
$$

- For the prior variance $\Sigma_0 = \frac{\sigma^2}{\lambda} I$, we get

$$
\hat{w} = \mu_P = \left(X^T X + \lambda I\right)^{-1} X^T y,
$$

which is of course the ridge regression solution.

9. Concept check questions Bayesian Method and Regression
10. Bayesian Methods

1 Gaussian Regression

This Gaussian regression model is one example of how we can find y from two approaches of 1) Frequentist and 2) Bayesian method.

Bayes conditional probablity model

formula

$$p(\theta \mid D, x) \propto \underbrace{p(D \mid \theta, x)}_{= L_D(\theta) \text{ likelihood}} \; \underbrace{p(\theta)}_{\text{prior}}$$

Gaussian Regression model

$$w \sim N(0, \Sigma_0)$$

$$y_i \mid x, w \overset{i.i.d}{\sim} N(w^T x_i, \sigma^2).$$

the conditional distributo of the random var y given x is gaussian w/ mean $w^T x$ & var σ^2.

$$p(w \mid D, x) \propto \underbrace{p(D \mid w, x)} \; \underbrace{p(w)}.$$

$$= \prod_{i=1}^{n} p(y_i \mid w, x) \quad p(w)$$

$$= \left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \sigma^2}} \cdot \exp\left(\frac{-1}{2\sigma^2} (y_i - w^T x_i)^2 \right) \right] \cdot$$

$$\left[(2\pi)^{-d/2} \det(\Sigma_0)^{-1/2} \exp\left(-\frac{1}{2} (w)^T \Sigma_0^{-1} (w) \right) \right]$$

$$(w - \mu_0)^T \Sigma_0^{-1} (w - \mu_0) = (w)^T \Sigma_0^{-1} (w)$$

$$\sim N(\mu_p, \Sigma_p)$$

for $\mu_p = (X^T X + \sigma^2 \Sigma_0^{-1})^{-1} X^T y$

$\Sigma_p = (\sigma^{-2} X^T X + \Sigma_0^{-1})^{-1}$

$$Y_{new} \mid x_{new}, D \sim N(\mu_{new}, \sigma^2_{new})$$

for $\mu_{new} = \mu_p^T x_{new}$

$\sigma^2_{new} = x^T_{new} \Sigma_p x_{new} + \sigma^2$

Posterior predictive ~~functio~~ distributio funct

b/c: $p(y_{new} \mid x_{new}, D) = \int p(y_{new} \mid x_{new}, w)\, p(w \mid D)\, \partial w$

predict the distribut° of Y for a given X

Goal: find $P(y_{new} \mid x_{new}, D)$.

→ approach 1 : find a point estimate for $w \xrightarrow{\text{frequentist}}$ MLE

approach 2 : unknown w is a variable, producing a distr on $w \in \mathbb{R}^d$ called the posterior distributo

⇒ get the distribut° for $Y \mid x$ by $\boxed{\text{integraty out } w}$

⟶ posterior predictive function

Find Y_{new} from 2 approaches — Frequentist / Bayesian

① Frequentist approach $\rightarrow$ MLE to find w^*

$$Y \mid x \sim N(w^T x, \sigma^2). \quad y$$

$$P_w(y_i \mid x_i) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(\frac{-1}{2\sigma^2}(y_i - w^T x_i)^2\right) \rightarrow \text{gets } P(Y \mid x_{new})$$

$$\Rightarrow y = \int pdf$$

$$L = P_w(D) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2\sigma^2}(y_i - w^T x_i)^2\right).$$

$$\log L = \sum_{i=1}^{n}\left[\log \frac{1}{\sigma\sqrt{2\pi}} + \left(\frac{-1}{2\sigma^2}(y_i - w^T x_i)^2\right)\right]$$

$$= n\log\frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w^T x_i)^2$$

Need to find $w^* = \underset{w}{\operatorname{argmax}} \log(P_w(D)) = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - w^T x_i)^2$

$$\frac{\partial P_w(D)}{\partial w} = -\frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} 2(y_i - w^T x_i)(-x_i)$$

$$= \frac{+1}{\sigma^2}\sum_{i=1}^{n}(y_i - w^T x_i)x_i$$

$$= 0$$

$$\Leftrightarrow \sum_{i=1}^{n}(y_i - w^T x_i)x_i = 0.$$

$\longleftarrow$ Try to remove Σ.

$$A = \sum_{i=1}^{n} (y_i - w^T x_i)^2 = (y - Xw)^T (y - Xw) \quad \text{for } X = \begin{bmatrix} - x_1 - \\ \vdots \\ - x_n - \end{bmatrix}$$

$$y = (y_1, .. y_n)^T.$$

$$= y^T y - \underline{2 w^T X^T y} + w^T X^T X w.$$

$$\frac{\partial c x}{\partial x} = c^T \qquad \frac{\partial x^T(Ax)}{\partial x} = (A^T + A)x.$$

$$\frac{\partial A}{\partial w} = -2 X^T y + \left[(X^T X)^T + (X^T X) \right] w$$

$$= -2 X^T y + 2 X^T X w$$

$$= 0$$

$$\Rightarrow X^T X w = X^T y.$$

$$\Rightarrow \boxed{w^* = (X^T X)^{-1} X^T y}$$

! Note $X^T X$ may not be invertible!!!

2. Bayesian Approach

w is treated as a var.

$\Rightarrow p(w)$ (prior distributⁿ of w) $\sim N(0, \Sigma)$.

$$\left.\begin{aligned} f(x_i) &= w^T x_i \\ \varepsilon_i &\sim N(0, \sigma^2) \\ y_i &= f(x_i) + \varepsilon_i \end{aligned}\right] \Rightarrow y_i | x_i \sim N(w^T x_i, \sigma^2).$$

In the matrix form:

$$X = \begin{bmatrix} - x_1 - \\ \vdots \\ - x_n - \end{bmatrix}, \quad y = (y_1, \dots, y_n)^T$$

$\Rightarrow \quad w \sim N(0, \Sigma)$.

$$f(X) = Xw$$

$$\varepsilon \sim N(0, \sigma^2 I)$$

$$I = \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{n \times n}$$

$$Y = f(X) + \varepsilon.$$

$\Rightarrow Y | X, w \sim N(Xw, \sigma^2 I) \Rightarrow \cancel{p(y|X,w) \propto \exp(\tfrac{-1}{2\sigma^2}}$

Posterior: $$p(w|D) = \frac{p(w \cap D)}{p(D)} = \frac{p(D|w)\,p(w)}{p(D)}$$

$$\begin{aligned} \Rightarrow p(w|D) &\propto p(D|w)\,p(w) \\ &= p(y|X,w)\,p(w) \\ &= N(y; Xw, \sigma^2 I)\,N(w; 0, \Sigma) \\ &\propto \exp\left(\frac{-1}{2\sigma^2}(Y - Xw)^T(Y - Xw)\right) \cdot \exp\left(-\frac{1}{2} w^T \Sigma^{-1} w\right) \\ &= \exp\left(\frac{-1}{2}\left[\frac{1}{\sigma^2}(Y - Xw)^T(Y - Xw) + w^T \Sigma^{-1} w\right]\right). \end{aligned}$$

Note $y_i \mid x_i, w \sim N(w^T x_i, \sigma^2)$

$$\Rightarrow p(y_i \mid x_i, w) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(\frac{-1}{2\sigma^2}(y_i - w^T x_i)^2\right).$$

We know: $P(D \mid X, w) = \prod_{i=1}^{n} p(y_i, x_i; w)$.

$$\propto \prod_{i=1}^{n} \exp\left(\frac{-1}{2\sigma^2}(y_i - w^T x_i)^2\right)$$

$$= \exp\left(\frac{-1}{2\sigma^2} \cdot \underbrace{\sum_{i=1}^{n} (y_i - w^T x_i)^2}\right).$$

$$= \exp\left(\frac{-1}{2\sigma^2} \cdot (Y - Xw)^T (Y - Xw)\right).$$

$\circ)$ ~~p(w|D)~~ =

$$\frac{1}{\sigma^2}(y - Xw)^T(y - Xw) + w^T \Sigma^{-1} w = \frac{1}{\sigma^2}\left[y^T y - 2w^T X^T y + w^T X^T X w\right] + w^T \Sigma^{-1} w$$

$$= w^T \underbrace{\left(\frac{1}{\sigma^2} X^T X + \Sigma^{-1}\right)}_{M.} w - 2\frac{1}{\sigma^2}(y^T X w) + \frac{1}{\sigma^2} y^T y.$$

$$\left(M = \frac{1}{\sigma^2} X^T X + \Sigma^{-1}, \quad b = \frac{1}{\sigma^2} X^T y.\right)$$

$$= w^T M w - 2b^T w + \frac{1}{\sigma^2} y^T y$$