Linear Regression

Regression problem

Given:

- observations
  $X = \{x_1, x_2, \dots, x_N\}$, $x_i \in \mathbb{R}^D$
- targets
  $y = \{y_1, y_2, \dots, y_N\}$, $y_i \in \mathbb{R}$

Find:

- mapping $f(\cdot)$ from inputs to targets
  $y_i \approx f(x_i)$

$$y_i = f(x_i) + \underbrace{\varepsilon_i}_{\text{noise}}, \quad \varepsilon_i \sim N(0, \beta^{-1})$$

$f(x)$ → linear function

$$f_w(x_i) = w_0 + w_1 x_{i1} + w_2 x_{i2} + \dots + w_D x_{iD}$$

$$= w_0 + w^T x_i$$

w_0: bias (offset); w: weight; x_i: features of x_i

→ absorb bias term

$$\tilde{x} = (1, x_1, \dots, x_D)^T$$
$$\tilde{w} = (w_0, w_1, \dots, w_D)^T$$

$$f_w(x) = \tilde{w}^T \tilde{x} \Rightarrow \boxed{f_w(x) = w^T x}$$

Loss function

Measures the "misfit" or error between our model (parametrized by w) and observed data $D = \{(x_i, y_i)\}_{i=1}^N$

- Least squares (LS)

$$E_{LS}(w) = \frac{1}{2} \sum_{i=1}^{N} (f_w(x_i) - y_i)^2 = \frac{1}{2} \sum_{i=1}^{N} (w^T x_i - y_i)^2$$

Objective: find optimal weight vector w^* that minimizes the error

$$w^* = \arg\min_w E_{LS}(w) = \arg\min_w \frac{1}{2} \sum_{i=1}^{N} (x_i^T w - y_i)^2 = \arg\min_w \frac{1}{2} (Xw - y)^T (Xw - y)$$

To find the minimum of the loss $E(w)$, compute the gradient $\nabla_w E(w)$:

$$\nabla_w E_{LS}(w) = \nabla_w \frac{1}{2} (Xw - y)^T (Xw - y)$$

$$= \nabla_w \frac{1}{2} (w^T X^T X w - 2 w^T X^T y + y^T y$$

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

Then set it to zero and solve for w to obtain the minimizer

$$X^T X w - X^T y \stackrel{!}{=} 0$$

$$w^* = \underbrace{(X^T X)^{-1} X^T}_{= X^\dagger \text{ (Moore-Penrose pseudo inverse of } X)} y$$

Nonlinear dependency in data

Solution: basis functions (polynomials, gaussians, logistic sigmoid)

$$f_w(x) = w_0 + \sum_{j=1}^{M} w_j x^j \quad \text{more generally} \quad f_w(x) = w_0 + \sum_{j=1}^{M} w_j \phi_j(x)$$

Defining $\phi_0 = 1$ $\quad f_w(x) = w^T \phi(x)$

The function f is still linear in w (despite not being linear in x)

$$f_w(x) = w_0 + \sum_{j=1}^{M} w_j \phi_j(x) = w^T \phi(x)$$

$$E_{LS}(w) = \frac{1}{2} \sum^{N} (w^T \phi(x_i) - y_i)^2 = \frac{1}{2} (\Phi w - y)^T (\Phi w - y)$$

$$\Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_M(x_1) \\ \vdots & & & \vdots \\ \phi_0(x_N) & \phi_1(x_N) & \cdots & \phi_M(x_N) \end{pmatrix} \in \mathbb{R}^{N \times (M+1)}$$

Optimal weights w^* : $\quad w^* = (\Phi^T \Phi)^{-1} \Phi^T y = \Phi^\dagger y$

To control overfitting → we introduce regularization

$$E_{ridge}(w) = \underbrace{\frac{1}{2}\sum^{N}\left[w^T\phi(x_i) - y_i\right]^2}_{\text{large values}} + \underbrace{\frac{\lambda}{2}\|w\|_2^2}_{\text{low values}}$$

- $\|w\|_2^2 \equiv w^Tw = w_0^2 + w_1^2 + \dots + w_M^2$ - squared L2 norm of w
- λ → regularization strength

The error of an estimator can be decomposed into 2 parts:

- Bias → expected error due to model mismatch
- Variance → variation due to randomness in training data

High bias → the model is too rigid to fit the underlying data distribution (regularization strength λ is too high)

High variance → the model is too flexible, and therefore captures noise in the data OVERFITTING (high capacity, memorizes data and λ too low)

WHAT IS IDEAL → low bias and low variance

The weights w_i can be interpreted as the strength of the (linear) relationship between feature x_i and y

But correlation does not imply causation

Probabilistic Linear Regression

$$y_i = f_w(x_i) + \underbrace{\varepsilon_i}_{\text{(noise)}} \qquad \varepsilon_i \sim N(0, \beta^{-1})$$

$$y_i \sim N(f_w(x_i), \beta^{-1})$$

maximum likelihood

of a single sample $\rightarrow$ $p(y_i \mid f_w(x_i), \beta) = N(y_i \mid f_w(x_i), \beta^{-1})$

entire dataset $D = \{X, y\}$ $\rightarrow$ $p(y \mid X, w, \beta) = \prod_{i=1}^{N} p(y_i \mid f_w(x_i), \beta)$

$$w_{ML}, \beta_{ML} = \underset{w,\beta}{\operatorname{argmax}}\ p(y \mid X, w, \beta) = \underset{w,\beta}{\operatorname{argmax}}\ \ln p(y \mid X, w, \beta) =$$

$$\underset{w,\beta}{\operatorname{argmin}} - \ln p(y \mid X, w, \beta)$$

maximum likelihood error function

$$E_{ML}(w, \beta) = -\ln p(y \mid X, w, \beta)$$

$$E_{ML}(w, \beta) = -\ln p(y \mid X, w, \beta) = -\ln \left[\prod_{i=1}^{N} p(y_i \mid f_w(x_i), \beta) \right]$$

$$= -\ln \left[\prod_{i=1}^{N} \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2} (w^T \phi(x_i) - y_i)^2 \right) \right]$$

$$= -\sum_{i=1}^{N} \ln \left[\sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{\beta}{2} (w^T \phi(x_i) - y_i)^2 \right) \right]$$

$$= \frac{\beta}{2} \sum_{i=1}^{N} (w^T \phi(x_i) - y_i)^2 - \frac{N}{2} \ln \beta + \frac{N}{2} \ln 2\pi$$

Optimizing log-likelihood

$$w_{ML} = \underset{w}{\operatorname{argmin}}\ E_{ML}(w, \beta) = \underset{w}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{N} (w^T \phi(x_i) - y_i)^2 = \underset{w}{\operatorname{argmin}}\ E_{LS}(w)$$

Maximizing the likelihood is equivalent to minimizing the least squares error function

$$\beta_{ML} = \arg\min_{\beta} E_{ML}(w_{ML}, \beta)$$

$$= \arg\min_{\beta} \left[\frac{\beta}{2} \sum_{i=1}^{N} (w_{ML}^T \phi(x_i) - y_i)^2 - \frac{N}{2} \ln \beta + \frac{N}{2} \ln 2\pi \right]$$

take derivative β and set it to 0

$$\frac{\partial}{\partial \beta} E_{ML}(w_{ML}, \beta) = \frac{1}{2} \sum_{i=1}^{N} (w_{ML}^T \phi(x_i) - y_i)^2 - \frac{N}{2\beta} \stackrel{!}{=} 0$$

Solving for β $\quad \frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{i=1}^{N} (w_{ML}^T \phi(x_i) - y_i)^2$

Posterior distribution

likelihood · prior

$$p(w \mid X, y, \beta, \cdot) = \frac{p(y \mid X, w, \beta) \cdot p(w \mid \cdot)}{p(y \mid X, \beta, \cdot)} \propto p(y \mid X, w, \beta) \cdot p(w \mid \cdot)$$

normalizing constant

Prior for w

$$p(w \mid \alpha) = \mathcal{N}(w \mid 0, \alpha^{-1} I) = \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left(-\frac{\alpha}{2} w^T w\right)$$

$\alpha \rightarrow$ precision of the distribution

$M \rightarrow$ n° of elements in the vector w

Maximum a posteriori (MAP)

$$w_{MAP} = \arg\max_{w} p(w \mid X, y, \alpha, \beta) = \arg\max_{w} \ln p(y \mid X, w, \beta) + \ln p(w \mid \alpha) - \underbrace{\ln p(y \mid X, \beta, \alpha)}_{\text{const}}$$

$$= \arg\min_{w} - \ln p(y \mid X, w, \beta) - \ln p(w \mid \alpha)$$

$$E_{MAP}(w) = -\ln p(y \mid X, w, \beta) - \ln p(w \mid \alpha) + \text{const}$$

$$= \frac{\beta}{2} \sum_{i=1}^{N} (w^T \phi(x_i) - y_i)^2 + \frac{\lambda}{2} \|w\|_2^2 + \text{const} \quad \propto E_{ridge}(w) + \text{const}$$

MAP estimation with Gaussian prior is equivalent to ridge regression

Full Bayesian approach

The full posterior distribution $p(w|D) \propto p(y|X,w,\beta) \cdot p(w|\alpha)$

Since both likelihood and prior are Gaussian, the posterior is as well

$$p(w|D) = N(w|\mu, \Sigma)$$

$$\mu = \beta \Sigma \Phi^T y \quad // \quad \Sigma^{-1} = \alpha I + \beta \Phi^T \Phi$$

$w_{MAP} = \mu$ for $N=0$, posterior = prior

$\alpha \to 0 \Rightarrow w_{MAP} \to w_{ML}$

Predictive distribution

$\hat{y}_{new}$ for new data x_{new}

- Maximum likelihood: w_{ML} and β_{ML}

$$p(\hat{y}_{new} | x_{new}, w_{ML}, \beta_{ML}) = N(\hat{y}_{new} | w_{ML}^T \phi(x_{new}), \beta_{ML}^{-1})$$

- Maximum a posteriori: w_{MAP}

$$p(\hat{y}_{new} | x_{new}, w_{MAP}, \beta) = N(\hat{y}_{new} | w_{MAP}^T \phi(x_{new}), \beta^{-1})$$

posterior proclictive distribution

$$p(\hat{y}_{new} | x_{new}, D) = N(\hat{y}_{new} | \mu^T \phi(x_{new}), \beta^{-1} + \phi(x_{new}) \Sigma \phi(x_{new}))$$