04-Linear Regression (Probabilistic View)

Probabilistic assumption

Assume the true value t (groundtruth) and the input $\boldsymbol{x}$ satisfies the following equation

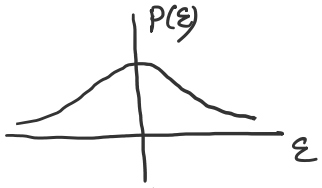
$$t = \boldsymbol{w}^T\boldsymbol{x} + \varepsilon$$

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

That is to say, t is indeed a linear/affine transformation of $\boldsymbol{x}$, except that some Gaussian noise affects t in an additive way.

The normal distribution $N(0,\sigma^2)$:

$$p(\varepsilon) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{\varepsilon}{\sigma}\right)^2\right\}$$



σ controls how fat the distribution is.

It is required to know Gaussian is quadratic in exponential decay. Memorizing the normalizing factor is not required in exams.

Since we know $\varepsilon = t - \boldsymbol{w}^T\boldsymbol{x}$

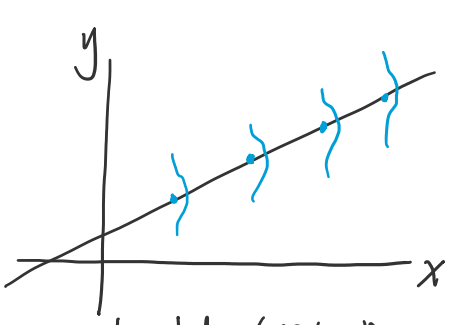
$$p(\varepsilon) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{t-\boldsymbol{w}^T\boldsymbol{x}}{\sigma}\right)^2\right\}$$

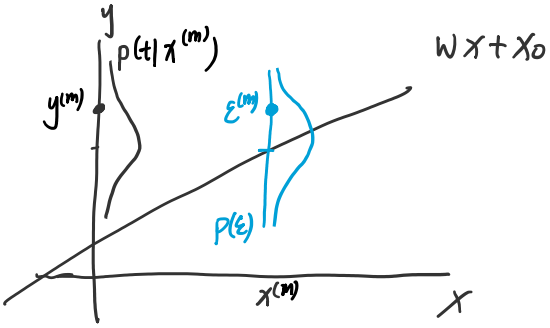
↑ The probability density of ε taking some value.

⇕

The probability density of t taking some corresponding value given $\boldsymbol{x}$,

$$p(t|\boldsymbol{x};\boldsymbol{w}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{t-\boldsymbol{w}^T\boldsymbol{x}}{\sigma}\right)^2\right\}$$





Knowledge/assumption A data sample

Maximum likelihood estimation

Suppose we have a set of data $\mathcal{D}$

We also have some probabilistic modeling of data

$p(\mathcal{D}; \boldsymbol{\theta})$

where $\boldsymbol{\theta}$ is the parameter.

Likelihood of parameters is yet another terminology of data probability

$\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) \triangleq p(\mathcal{D}; \boldsymbol{\theta})$

Maximum likelihood estimation (MLE) maximizes the likelihood of parameters, which is the probability of data.

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta})$$

[log is monotonically increasing]

log of likelihood is known as log-likelihood, denoted by $\ell(\theta; \mathcal{D})$

In our probabilistic assumption for linear regression, we consider the probability of all labels $t^{(i)}$ given $\boldsymbol{x}^{(i)}$, parameterized by model parameters.

$$\log \mathcal{L}(\boldsymbol{w}) = \log p(t^{(1)}, t^{(2)}, \cdots, t^{(M)} \mid \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots \boldsymbol{x}^{(M)})$$

[y's are RVs, x's not]

$$= \log \prod_{m=1}^{M} p(t^{(m)} \mid \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots, \boldsymbol{x}^{(M)})$$

[data iid]

$$= \log \prod_{m=1}^{M} p(t^{(m)} \mid \boldsymbol{x}^{(m)})$$

[$y^{(i)}$ only depends on $\boldsymbol{x}^{(i)}$]

$$= \sum_{m=1}^{M} \log p(t^{(m)} \mid \boldsymbol{x}^{(m)})$$

[$\log \prod = \sum \log$]

$$= \sum_{m=1}^{M} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{t^{(m)} - \boldsymbol{w}^T \boldsymbol{x}^{(m)}}{\sigma}\right)^2\right\}$$

$$= \sum_{m=1}^{M} \left[\log \frac{1}{\sqrt{2\pi}\sigma} + \log \exp\left\{-\frac{1}{2}\left(\frac{t^{(m)} - \boldsymbol{w}^T \boldsymbol{x}^{(m)}}{\sigma}\right)^2\right\}\right]$$

$$= -M \cdot \log(\sqrt{2\pi}\sigma) - \frac{1}{2} \sum_{m=1}^{M} \left(\frac{t^{(m)} - \boldsymbol{w}^T \boldsymbol{x}^{(m)}}{\sigma}\right)^2$$

Thus, maximize $\log \mathcal{L}(\boldsymbol{w})$

$$\Leftrightarrow \quad \text{minimize}_{\boldsymbol{w}} \quad \sum_{m=1}^{M} \left(t^{(m)} - \boldsymbol{w}^T \boldsymbol{x}^{(m)} \right)^2$$

$$\Leftrightarrow \quad \text{minimize}_{\boldsymbol{w}} \quad \frac{1}{2M} \sum_{m=1}^{M} \left(t^{(m)} - \boldsymbol{w}^T \boldsymbol{x}^{(m)} \right)^2$$

This is exactly MSE.

Best Linear Unbiased Estimate (BLUE)

Theorem. Assume $t^{(i)}$ is generated by $t^{(i)} = \boldsymbol{\omega}^T \boldsymbol{x} + \epsilon^{(i)}$, where $\epsilon^{(i)}$ is zero-mean, uncorrelated, and with finite variance; $\boldsymbol{\omega}$ is unknown parameters. Then, least square error yields an unbiased estimate of $\boldsymbol{\omega}$.

Proof: The closed-form solution is $\hat{\boldsymbol{w}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{t}$

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{E}^{(m)}}[\hat{\boldsymbol{w}}] &= \mathop{\mathbb{E}}_{\mathcal{E}^{(m)}}\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{t} \right] \\
&= \mathop{\mathbb{E}}_{\mathcal{E}^{(m)}}\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1} \boldsymbol{X}^T (\boldsymbol{X}\boldsymbol{w} + \boldsymbol{\mathcal{E}}) \right] \qquad [\, \boldsymbol{\mathcal{E}} = (\mathcal{E}^{(1)}, \cdots \mathcal{E}^{(M)})^T \,] \\
&= \mathop{\mathbb{E}}_{\mathcal{E}^{(m)}}\left[\underbrace{(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})}_{=\boldsymbol{I}} \boldsymbol{w} + \underbrace{\mathop{\mathbb{E}}_{\mathcal{E}^{(m)}}\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1} \boldsymbol{X}^T \right]}_{=\boldsymbol{0}} \right] \\
&= \boldsymbol{w}
\end{aligned}$$

#

Theorem (Gauss-Markov). Least error regression yields the least variance in linear, unbiased estimators. (Here, a linear estimator means $\boldsymbol{w} = \boldsymbol{C}\boldsymbol{t}$ for some $\boldsymbol{C}$.)
https://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem #

Bias-Variance Tradeoff

- Limitation of the above analysis:
  In reality the assumption $t = \boldsymbol{w}^T \boldsymbol{x} + \mathcal{E}$ is usually not true, and we do not know the true function.

Now suppose $t = f(\boldsymbol{x}) + \mathcal{E}$ where $\mathcal{E} \sim p(\mathcal{E}|\boldsymbol{x})$ and $\mathop{\mathbb{E}}_{\mathcal{E} \sim p(\mathcal{E}|\boldsymbol{x})}[\mathcal{E}] = 0 \ \forall \boldsymbol{x}$

In other words, we can treat t as an R.V. and define $f(\boldsymbol{x}) = \mathop{\mathbb{E}}_{t \sim p(t|\boldsymbol{x})}[t]$

We consider the performance of a machine learning model $h(\boldsymbol{x})$ by the measure of success. In our case, the measure of success is modeled by MSE. We call how bad the model performs on a data sample $\boldsymbol{x}$ the error on $\boldsymbol{x}$, denoted by $E(\boldsymbol{x})$. In here, it happens to be the same as the training objective function. We call the expected loss of a model h (when data are repeatedly generated from the data distribution) error, too, denoted by $E(h)$. This is also known as the risk.

$$E(h) = \mathop{\mathbb{E}}_{\boldsymbol{x},\varepsilon}\left[\left(f(\boldsymbol{x}) + \varepsilon - h(\boldsymbol{x})\right)^2\right]$$

[Recall: $h(\boldsymbol{x})$ is the prediction, $f(\boldsymbol{x})$ is the "true" function, less noise]

$$= \mathop{\mathbb{E}}_{\boldsymbol{x},\varepsilon}\left[\left(f(\boldsymbol{x}) - h(\boldsymbol{x}) + \varepsilon\right)^2\right]$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{x},\varepsilon}\left[\left(f(\boldsymbol{x}) - h(\boldsymbol{x})\right)^2\right] + \mathop{\mathbb{E}}_{\boldsymbol{x},\varepsilon}\left[\left(f(\boldsymbol{x}) - h(\boldsymbol{x})\right)\cdot\varepsilon\right] + \mathop{\mathbb{E}}_{\boldsymbol{x},\varepsilon}\left[\varepsilon^2\right] \quad (\mathbb{E} \text{ is linear})$$

$= 0$, because $\mathop{\mathbb{E}}_{\varepsilon}[\varepsilon] = 0$

$$= \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\left(f(\boldsymbol{x}) - h(\boldsymbol{x})\right)^2\right] + \mathop{\mathbb{E}}_{\substack{\boldsymbol{x}\sim p(\boldsymbol{x}) \\ \varepsilon\sim p(\varepsilon|\boldsymbol{x})}}\left[\varepsilon^2\right] \quad (1)$$

Notice that $h(\boldsymbol{x})$, in fact, depends on the training data $\mathcal{D}$. We denote it by $h_{\mathcal{D}}(\boldsymbol{x})$. The first term can be further decomposed as

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\left(f(\boldsymbol{x}) - h_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] = \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\left\{f(\boldsymbol{x}) - \mathop{\mathbb{E}}_{\mathcal{D}'}[h_{\mathcal{D}'}(\boldsymbol{x})] + \mathop{\mathbb{E}}_{\mathcal{D}'}[h_{\mathcal{D}'}(\boldsymbol{x})] - h_{\mathcal{D}}(\boldsymbol{x})\right\}^2\right]$$

▷ Did nothing, only introduced $\pm \mathop{\mathbb{E}}_{\mathcal{D}'}[h_{\mathcal{D}'}(\boldsymbol{x})]$

$$= \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\left\{f(\boldsymbol{x}) - \mathop{\mathbb{E}}_{\mathcal{D}'}[h_{\mathcal{D}'}(\boldsymbol{x})]\right\}^2\right] + \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\left\{\mathop{\mathbb{E}}_{\mathcal{D}'}[h_{\mathcal{D}'}(\boldsymbol{x})] - h_{\mathcal{D}}(\boldsymbol{x})\right\}^2\right]$$

$$+ \mathop{\mathbb{E}}_{\boldsymbol{x}}\left[2\left(f(\boldsymbol{x}) - \mathop{\mathbb{E}}_{\mathcal{D}'}[h_{\mathcal{D}'}(\boldsymbol{x})]\right)\cdot\left(\mathop{\mathbb{E}}_{\mathcal{D}'}[h_{\mathcal{D}'}(\boldsymbol{x})] - h_{\mathcal{D}}(\boldsymbol{x})\right)\right] \quad (2)$$

We further take expectation of data $\mathcal{D}$ (assuming data can be iid generated from a repeated trial).

In this case $\mathop{\mathbb{E}}_{\mathcal{D}}[\text{last term above}] = 0$ because $\mathop{\mathbb{E}}_{\mathcal{D}}\left[\mathop{\mathbb{E}}_{\mathcal{D}'}[h_{\mathcal{D}'}(\boldsymbol{x})] - h_{\mathcal{D}}(\boldsymbol{x})\right] = 0$

Combining (1) and (2), we have

$$\mathop{\mathbb{E}}_{\mathcal{D}}[R(h_{\mathcal{D}})] = \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\left\{f(\boldsymbol{x}) - \mathop{\mathbb{E}}_{\mathcal{D}'}[h_{\mathcal{D}'}(\boldsymbol{x})]\right\}^2\right]}_{\text{bias}^2} + \underbrace{\mathop{\mathbb{E}}_{\mathcal{D},\boldsymbol{x}}\left[\left\{\mathop{\mathbb{E}}_{\mathcal{D}'}[h_{\mathcal{D}'}(\boldsymbol{x})] - h_{\mathcal{D}}(\boldsymbol{x})\right\}^2\right]}_{\text{variance}} + \underbrace{\mathop{\mathbb{E}}_{\substack{\boldsymbol{x}\sim p(\boldsymbol{x}) \\ \varepsilon\sim p(\varepsilon|\boldsymbol{x})}}\left[\varepsilon^2\right]}_{\text{noise}}$$

Interpretation: The expected error (said in terms of training data $\mathcal{D}$ repeated drawn from a repeatable trial and test data $\boldsymbol{x}$, ϵ following test distribution) is decomposed of

Expected error = Bias^2 + Variance + Noise

- **Bias:** The difference between the true function and the expected learned function
- **Variance**: The square difference between the learned function h from this training data $\mathcal{D}$ and the expected learned function
- **Noise:** Variance of noise, which is intrinsic to the dataset and cannot be optimized

Reducing bias:

- Increasing the hypothesis class (containing the true function is a necessary condition for being unbiased)

Reducing variance:

- Reducing the hypothesis class (fewer functions to choose from)

Example: Suppose the truth function is nth-degree polynomial

$$t = w_1 x + w_2 x^2 + \cdots + w_n x^n$$

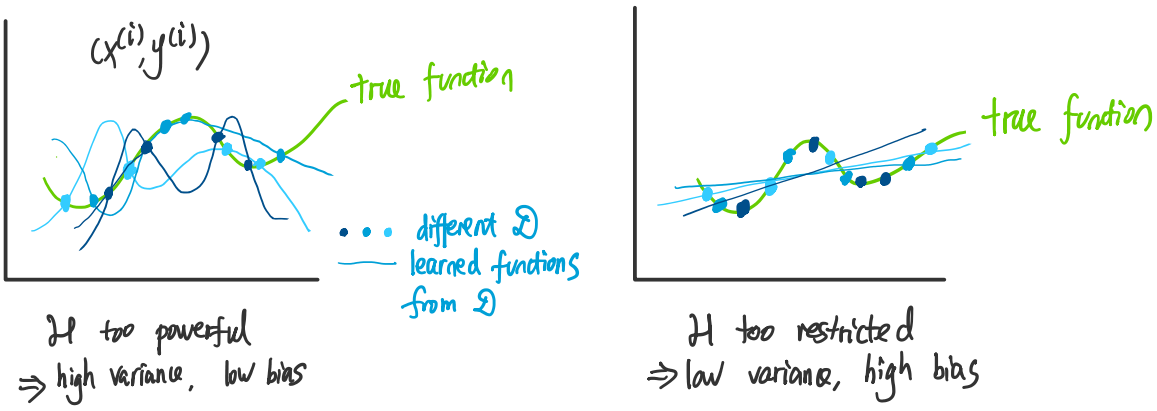
We can introduce polynomial basis as the feature set.

$$x_1 = x, \quad x_2 = x^2, \quad \cdots, \quad x_d = x^d, \quad \text{and} \quad x_0 = 1 \text{ as usually.}$$

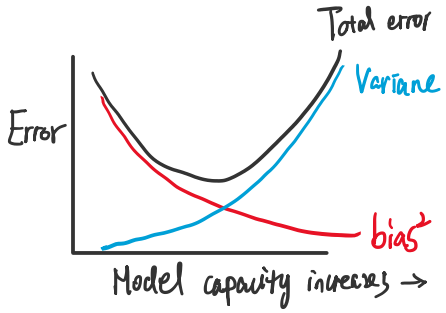
In this case, the least square estimate is still unbiased if $d \geq n$ (as it has the same format as linear regression). The resulting model is still linear in $\boldsymbol{w}$ and $(x_1, \ldots, x_d)^T$, although not linear in $\boldsymbol{x}$.

Unfortunately, we do not know the true order of polynomial n.

- If d is too large, the model is unbiased, but since we have too many functions to choose from, the estimator is of high variance.
- If, on the other hand, d is too small, the model is biased, but we have fewer functions to choose from and the estimator is of low variance.



An intuitive picture of bias-variance tradeoff



An approach to the bias-variance tradeoff

- We must state preference of $\mathcal{H}$ before training
- Instead of a highly restricted form of functions (e.g., linear with limited features), we
  - Allow a more powerful form of functions, but
  - State some other preferences, e.g.,
    - Among $w_1, w_2, \ldots, w_d$, only a few parameters could be non-zero, or
    - The weights $w_1, w_2, \ldots, w_d$ cannot be too large, etc.

Footnote on the terminologies

- In machine learning, people sometimes use *objective*, *loss*, *risk*, *error* interchangeably. They are indeed similar, but may have different emphasis.
- *Objective*, or *training objective*, is clear. It is the goal of training, which may or may not be the same function as loss/risk/error.
- *Loss* has two meanings.
  - When we say training loss, it usually means training objective.
  - In decision theory (where they don't care about training), *loss*

    - In decision theory (where they don't care about training), *loss* means the measure of success for a model h evaluated on a particular data point $L(x, h)$. We use *Error* in this course. The lower, the better.
- *Risk* is usually said by referring to the expected loss/error when data are repeated generated, so it's also called the *frequentist risk*. It is function of your predictor h and the true model f, denoted by $R(h, f)$. It should be emphasized that the expecation is said in terms of data being repeated generated from the data distribution. So this terminology actual makes some sense, because we have uncertainty of the world (data), and thus, risk.

Reference: Chap 1.3, James O. Berger, *Statistical Decision Theory and Bayesian Analysis*. https://link.springer.com/book/10.1007/978-1-4757-4286-2