

Notes On Machine Learning

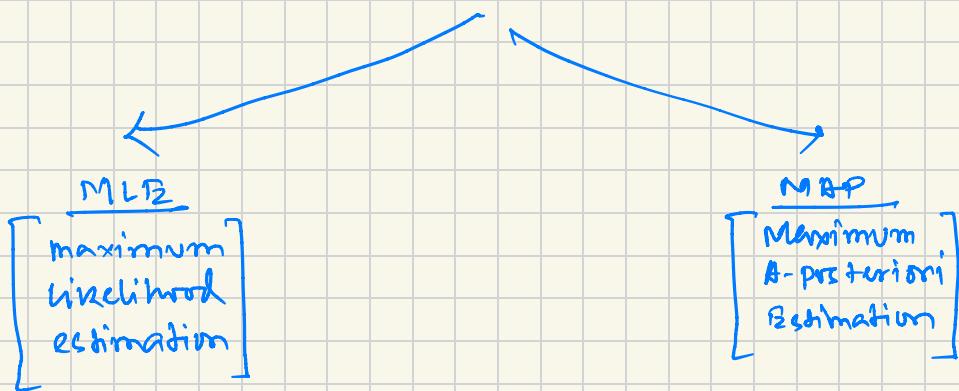
Parameter Estimation

- ** given a parameter value of a distribution how the density/distribution function looks like.
- ** Sampling: If x follows a distribution, how can we obtain different values of x will take.
- ** Parameter estimation: Given different values of x , how can we obtain the parameters of the underlying function.
 - * Given data, statistics/ML aim to learn parameters of the underlying distribution from data.
 - Gaussian Mixture Model / Probabilistic graphical model / linear regression / logistic regression.
 - * Sampling is essential in probabilistic machine learning & Bayesian statistics.
 - Latent dirichlet allocation / Bayesian process / probabilistic graphical model.
 - Markov chain model (MCMC - Markov chain Monte Carlo).
- * Any statistic used to estimate the value of an unknown parameter θ is called an estimator of θ .
 - Ex - μ & σ for normal distribution
 λ for precision.

The strategy is to choose the parameter in such a way that maximize the likelihood between the prediction & the label or minimize the loss between the prediction & true label.

Once you define loss function only need to optimize the loss function with the parameter.

Two ways to defining loss function



① LSE (Least Square Error)

This is special form of MLE. we may define the loss function as:

$$\hat{w}_{LSE} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^N (\hat{y}_i - y_i)^2$$

$$= \underset{w}{\operatorname{argmin}} \sum_{i=1}^N (x_i w_i - y_i)^2$$

$$= \underset{w}{\operatorname{argmin}} (x w - y)^2$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{bmatrix}$$

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \\ x_m & \cdots & \cdots & x_{mp} \end{bmatrix}$$

This is closed form of above

Now we need to choose w that minimizes

$$(x w - y)^2$$

So, Loss function

$$L(w) = \gamma_2 \sum_{i=1}^N (x_i w_i - y_i)^2 \\ = \gamma_2 (xw - y)^T$$

$$\frac{\partial L(w)}{\partial w} \Rightarrow (xw - y)x^T = 0 \\ \Rightarrow x^T x w - x^T y = 0 \\ \Rightarrow w = (x^T x)^{-1} x^T y$$

Thus, $\hat{w} = (x^T x)^{-1} x^T y$ gives least square estimation for our parameter w .

② Maximum Likelihood Estimation (MLE)

MLE to find the parameters that maximize the likelihood between y & \hat{y} .

We can define the loss function of MLE as

$$\hat{\theta}_{MLE} = \operatorname{argmax}_\theta P(D|\theta)$$

this represents probability on the dataset D, calculated by the parameter θ .

Parameter to estimate which is similar to w

D is the dataset

This is similar to $P(\hat{x}|y) = P(x.w|y)$

Ex: MLE for Bernoulli parameter!

$$\text{prob} = P(x) = \begin{cases} 1-p & \text{for } x=0 \\ p & \text{for } x=1 \end{cases}$$

It is also written as $p^n(1-p)^{n-1}$ as a closed form.

$$f(x_1, x_2, \dots, x_n | p) = p(x=x_1, x=x_2, x=x_3, \dots, x=x_n | p)$$

$$\Rightarrow p(x=x_1 | p) p(x=x_2 | p) \cdots p(x=x_n | p)$$

$$\Rightarrow p^{x_1} (1-p)^{1-x_1} \cdot p^{x_2} (1-p)^{1-x_2} \cdots p^{x_n} (1-p)^{1-x_n}$$

$$\Rightarrow p^{\sum_{i=1}^n x_i} (1-p)^{\sum_{i=1}^n 1 - n}$$

taking log both side to find log likelihood.

$$\log f(x_1, x_2, \dots, x_n | p) = \sum_{i=1}^n x_i \log p + \left(\sum_{i=1}^n 1 - n \right) \log (1-p)$$

$$\frac{\partial}{\partial p} (\log f(x_1, x_2, \dots, x_n | p)) = \frac{1}{p} \sum_{i=1}^n x_i + \frac{1}{1-p} \left(\sum_{i=1}^n 1 - n \right) = 0$$

$$\Rightarrow \frac{1}{p} \sum_{i=1}^n x_i + \frac{1}{1-p} \sum_{i=1}^n 1 - n = 0$$

$$\Rightarrow \sum_{i=1}^n x_i \left(\frac{1}{p} + \frac{1}{1-p} \right) = \frac{n}{1-p}$$

$$\Rightarrow \sum_{i=1}^n x_i \left(\frac{1-p+p}{p(1-p)} \right) = \frac{n}{1-p}$$

$$\Rightarrow p = \frac{1}{n} \sum_{i=1}^n x_i$$

MLB for normal Distribution:

$$P(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

take log likelihood,

$$\log P(x | \mu, \sigma^2) = \log \left[\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right]$$

$$L = \sum_{i=1}^n \log \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right)$$

$$L = \sum_{i=1}^n \log \left(\frac{1}{\sqrt{2\pi}\sigma} \right) + \sum_{i=1}^n w_i e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$L_2 = \sum_{i=1}^n w_i \left(\frac{1}{\sqrt{2\pi}\sigma} \right) + \sum_{i=1}^n -\frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial L}{\partial \mu} = \sum_{i=1}^n -\frac{1}{2\sigma^2} 2(x_i-\mu) (-1)$$

$$= \sum_{i=1}^n \frac{x_i - \mu}{\sigma^2} = 0 \Rightarrow$$

$$\sum_{i=1}^n x_i - n\mu = 0$$

$$\Rightarrow \boxed{\mu = \frac{1}{n} \sum_{i=1}^n x_i}$$

$$\frac{\partial L}{\partial \sigma} = \frac{\partial}{\partial \sigma} n \log(2\pi\sigma^2)^{1/2} + -\frac{1}{2} \sum_{i=1}^n (\mu - \mu) \frac{\partial}{\partial \sigma} (\sigma^{-2})$$

$$= -\frac{n}{2} \frac{1}{2\pi\sigma^4} * 2\pi\sigma^2 + -\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 - 2\sigma^{-3}$$

$$\Rightarrow -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu)^2 = 0$$

$$\Rightarrow \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 = \sigma^2$$

$$\Rightarrow \boxed{\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2}}$$

③ Maximum A-posteriori Estimation (MAP)

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} P(\theta | D) = \underset{\theta}{\operatorname{argmax}} \frac{P(D|\theta) P(\theta)}{P(D)}$$

$$\propto \underset{\theta}{\operatorname{argmax}} P(D|\theta) P(\theta)$$

Since $P(D)$ in the above expression doesn't depend on θ , we can just maximize the numerator $P(D|\theta) P(\theta)$.

$$\begin{aligned}\hat{\theta}_{MAP} &= \underset{\theta}{\operatorname{argmax}} P(D|\theta) P(\theta) \\ &= \underset{\theta}{\operatorname{argmax}} \left\{ \log(P(D|\theta)) + \log P(\theta) \right\} \\ &= \underset{\theta}{\operatorname{argmax}} \left\{ \sum_{i=1}^n \log P(D_i|\theta) + \log P(\theta) \right\}\end{aligned}$$

This is same as MLE with extra log-prior-distribution term.

MAP allows incorporating our prior knowledge about θ in its distribution.

* prior belief of p :

→ This prior belief of p will be 0 outside $[0,1]$

→ In most cases we want to choose a distribution for prior beliefs that peaks somewhere in the interval $[0,1]$.

□ Beta distribution:

Parameterized by two shape constants $\alpha \geq \beta$.

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

$$\frac{1}{B(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1}$$

$$B = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

where $\Gamma_n = (n-1)!$

1.1 Definition

Beta distribution is a type of statistical distribution, which has two free parameters. It is used as a prior distribution in Bayesian inference, due to the fact that it is the conjugate prior distribution for the binomial distribution, which means that the posterior distribution and the prior distribution are in the same family.

The probability distribution function (pdf) of the beta distribution is defined as,

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) + \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} = \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)} \quad (1)$$

where $\alpha > 0, \beta > 0, x \in [0, 1]$ and $\Gamma(\cdot)$ denotes the gamma function.

1.2 Application

Considering the classical Bernoulli problem (repeated coin flipping), after n trials, there are s successes (heads) and f failures (tails). Let a random variable x denote the success probability of each trial. The likelihood for parameters s and f given $x = p$ is the following binomial distribution,

$$L(s, f | x = p) = \binom{n}{s} x^s (1-x)^{n-s} \quad (2)$$

If belief about prior probability information is reasonably well approximated by a beta distribution,

$$P(x = p; \alpha, \beta) = \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)} \quad (3)$$

According to Bayes's theorem, the posterior probability is given by the product of the likelihood function and the prior probability normalised by the integral as follows,

$$\begin{aligned} P(x = p | s, f) &= \frac{L(s, f | x = p) P(x = p; \alpha, \beta)}{\int_0^1 L(s, f | x = p) P(x = p; \alpha, \beta) dx} \\ &= \frac{\binom{n}{s} x^s (1-x)^{n-s} x^{\alpha-1} (1-x)^{\beta-1}}{\int_0^1 \binom{n}{s} x^s (1-x)^{n-s} x^{\alpha-1} (1-x)^{\beta-1} dx} \\ &= \frac{x^{\alpha+s-1} (1-x)^{n-s+\beta-1}}{B(s+\alpha, n-s+\beta)} \end{aligned} \quad (4)$$

The posterior probability function is also a beta distribution (conjugate). It is convenient to compute. This is the main reason why we approximate the prior using a beta distribution.

* Derive a-posteriori estimation for the parameter θ :

$$\hat{P}_{MAP} = \left(\prod_{i=1}^n P_{\text{Bernoulli}}(x_i | \theta) \right) \cdot P_{\text{Beta}}(\theta | \alpha, \beta)$$

$$L = \sum_{i=1}^n \log P_{\text{Bernoulli}}(x_i | \theta) + \log P_{\text{Beta}}(\theta | \alpha, \beta)$$

$$= \sum_{i=1}^n \log \left[\theta^{x_i} * (1-\theta)^{x_i-1} \right] + \log \frac{\alpha^{\alpha-1} (1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

$$L = \sum_{i=1}^n \log \theta^{x_i} + \sum_{i=1}^n \log (1-\theta)^{x_i-1} + \log \theta^{\alpha-1} + \log (1-\theta)^{\beta-1} - \log B(\alpha, \beta)$$

$$L = \sum_{i=1}^n x_i \log \theta + \sum_{i=1}^n (x_i-1) \log (1-\theta) + (\alpha-1) \log \theta + (\beta-1) \log (1-\theta) - \log B(\alpha, \beta)$$

$$\frac{\partial L}{\partial \theta} = \sum_{i=1}^n \frac{x_i}{\theta} - \sum_{i=1}^n \frac{(x_i-1)}{1-\theta} + \frac{\alpha-1}{\theta} - \frac{\beta-1}{1-\theta} = 0$$

$$\Rightarrow \frac{1}{\theta} \sum_{i=1}^n x_i + \frac{1}{1-\theta} \sum_{i=1}^n x_i - \frac{n}{1-\theta} + \frac{\alpha-1}{\theta} - \frac{\beta-1}{1-\theta} = 0$$

$$\Rightarrow \frac{1}{\theta} \left[\sum_{i=1}^n x_i + \alpha - 1 \right] + \frac{1}{1-\theta} \left[\sum_{i=1}^n x_i - n - \beta + 1 \right] = 0$$

$$\Rightarrow (1-\theta) \left[\sum_{i=1}^n x_i + \alpha - 1 \right] + \theta \sum_{i=1}^n x_i - n\theta - \beta\theta + \theta = 0$$

$$\Rightarrow \sum_{i=1}^n x_i + \alpha - 1 - \theta \sum_{i=1}^n x_i - \alpha\theta + \theta + \theta \sum_{i=1}^n x_i - n\theta - \beta\theta + \theta = 0$$

$$\Rightarrow \sum_{i=1}^n x_i + \alpha - 1 + \theta [\alpha - \beta - n + 2] = 0$$

$$\Rightarrow \theta [n + \alpha + \beta - 2] = \sum_{i=1}^n x_i + \alpha - 1$$

$$\Rightarrow \theta_{MAP} = \frac{\sum_{i=1}^n x_i + \alpha - 1}{n + \alpha + \beta - 2}$$

* Apply MAP on multinomial distribution

The multinomial distribution is a common distribution for characterizing categorical variables. Suppose a random variable Z has k categories, we can code each category as an integer, leading to $Z \in \{1, 2, \dots, k\}$. Suppose that $P(Z = k) = p_k$. The parameter $\{p_1, \dots, p_k\}$ describes the entire distribution of k (with the constraint that $\sum_j p_j = 1$). Suppose we generate Z_1, \dots, Z_n IID from the above distributions and let

$$X_j = \sum_{i=1}^n I(Z_i = j) = \# \text{ of observations in the category } j.$$

Then the random vector $X = (X_1, \dots, X_k)$ is said to be from a multinomial distribution with parameter (n, p_1, \dots, p_k) . We often write

$$X \sim M_k(n; p_1, \dots, p_k)$$

to denote a multinomial distribution.

Example (pet lovers). The following is a hypothetical dataset about how many students prefer a particular animal as a pet. Each row (except the 'total') can be viewed as a random vector from a multinomial distribution. For instance, the first row $(18, 20, 6, 4, 2)$ can be viewed as a random draw from a multinomial distribution $M_5(n = 50; p_1, \dots, p_5)$. The second and the third row can be viewed as other random draws from the same distribution.

	cat	dog	rabbit	hamster	fish	total
Class 1	18	20	6	4	2	50
Class 2	15	15	10	5	5	50
Class 3	17	18	8	4	3	50

7.1 Properties of multinomial distribution

The PMF of a multinomial distribution has a simple closed-form. If $X \sim M_k(n; p_1, \dots, p_k)$, then

$$p(X = x) = p(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}.$$

The *multinomial coefficient* $\frac{n!}{x_1! x_2! \dots x_k!} = \binom{n}{x_1, x_2, \dots, x_k}$ is the number of possible ways to put n balls into k boxes. The famous *multinomial expansion* is

$$(a_1 + a_2 + \dots + a_k)^n = \sum_{x_i \geq 0, \sum_i x_i = n} \frac{n!}{x_1! x_2! \dots x_k!} a_1^{x_1} a_2^{x_2} \dots a_k^{x_k}.$$

This implies that $\sum_{x_i \geq 0, \sum_i x_i = n} p(X = x) = 1$.

$$L(\theta | D) \propto p_{\text{multinomial}}(x_1, x_2, \dots, x_N | \alpha_1, \alpha_2, \dots, \alpha_N) \cdot p_{\text{Dirichlet}}(\alpha_1, \alpha_2, \dots, \alpha_N | \alpha_1, \alpha_2, \dots, \alpha_N)$$

$$= \ln \prod_{i=1}^N \frac{\alpha_i^{x_i}}{\Gamma(x_i)} \times \left[\frac{\Gamma(\sum_{j=1}^N \alpha_j)}{\prod_{j=1}^N \Gamma(\alpha_j)} \right] \prod_{i=1}^N \alpha_i^{x_i - 1}$$

$$\begin{aligned} & \text{removing all constant} \\ & \prod_{i=2}^N \alpha_i^{x_i} \quad \prod_{i=1}^N \alpha_i^{x_i - 1} \end{aligned}$$

$$= \prod_{i=1}^n \theta_i^{x_i + \alpha_i - 1}$$

$$L = \log \left(\prod_{i=1}^n \theta_i^{x_i + \alpha_i - 1} \right) = \sum_{i=1}^n (x_i + \alpha_i - 1) \log \theta_i$$

we have constraint $\sum_{i=1}^n \theta_i = 1$

To impose this constraint we will apply lagrange multiplier λ to log-likelihood.

$$L = \sum_{i=1}^n (x_i + \alpha_i - 1) \log \theta_i - \lambda \left(\sum_{i=1}^n \theta_i - 1 \right) = 0$$

taking derivative on single θ_i cancels out other terms

$$\frac{\partial L}{\partial \theta_i} = \frac{x_i + \alpha_i - 1}{\theta_i} - \lambda = 0 \Rightarrow \theta_i = \frac{x_i + \alpha_i - 1}{\lambda}$$

$$\sum_{i=1}^n \theta_i = \frac{1}{\lambda} \sum_{i=1}^n x_i + \alpha_i - 1 = 1$$

$$\Rightarrow \lambda = \sum_{i=1}^n x_i + \alpha_i - 1$$

so, replacing λ 's value

$$\hat{\theta}_{i \text{ MAP}} = \frac{x_i + \alpha_i - 1}{\sum_{i=1}^n x_i + \alpha_i - 1}$$

Linear Regression

Learn a function which maps input to output

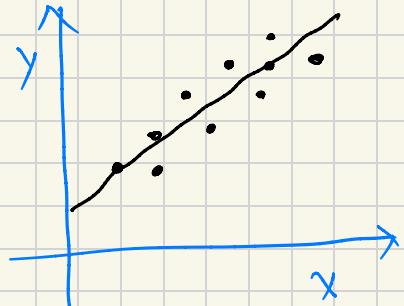
$$f: X \rightarrow Y$$

$$\hat{Y}_i = w_0 + w_1 x_i$$

$$= w_0 + w^T x$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{bmatrix} \quad p \times 1$$

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \\ x_m & \cdots & \cdots & x_{mp} \end{bmatrix} \quad m \times p$$



$$\text{Loss } L(w) = \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

$$L(w) = \frac{1}{2} \sum_{i=1}^N \{ y_i - (w^T x_i + w_0) \}^2$$

The goal is to minimize this loss function.

We can also write it as

$$\frac{\partial L(w)}{\partial w} = \frac{\partial}{\partial w} \left[\frac{1}{2} (w^T x - y)^2 \right]$$

$$\nabla E(w) \Rightarrow \frac{1}{2} \cdot 2 (w^T x - y) x^T = 0$$

$$2) \quad x^T x w = x^T y$$

$$\Rightarrow w = (x^T x)^{-1} x^T y$$

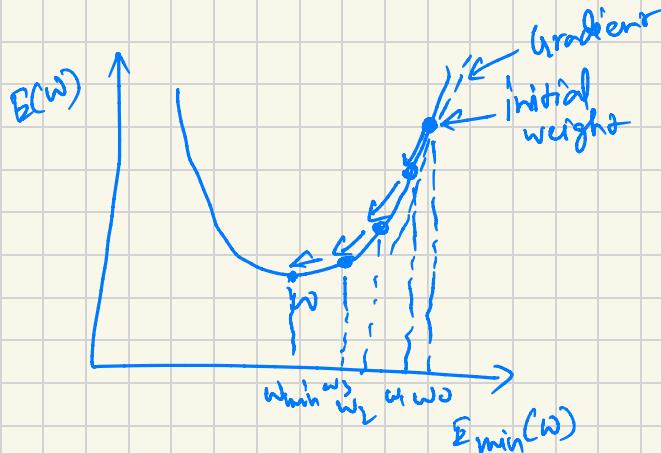
$\underbrace{x^T x}_{P \times N}$ $\underbrace{(x^T x)^{-1}}_{N \times P}$ $\underbrace{x^T y}_{P \times N}$ $\underbrace{y}_{N \times 1}$
 $\underbrace{P \times P}_{P \times 1}$ $\underbrace{P \times 1}_{P \times 1}$

~~* *~~ update weight

$$\omega^{t+1} = \omega^t - \eta \nabla E \rightarrow (x\omega - y)x^T$$

t = iteration number

η = learning rate parameter



$$w_1 = w_0 - \eta \nabla E$$

$$w_2 = w_1 - \eta \nabla E$$

$$w_3 = w_2 - \eta \nabla E$$

$$W_{\min} = w_3 - \eta \nabla E$$

~~* *~~ To apply gradient descent, the function needs to be a convex function.

→ This function will have only one global minimum.

→ It follows a property called Jensen's Inequality.

Least Squares Solution:

Write down three equations for the line $y = mx + c$ to go through $y=7$ at $x=-1$, $y=7$ at $x=1$ and $y=21$ at $x=2$. Find the least squares solution (c, m) ?

$$7 = -m + c$$

$$7 = m + c$$

$$21 = 2m + c$$

$$\begin{bmatrix} -1 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} m \\ c \end{bmatrix} = \begin{bmatrix} 7 \\ 7 \\ 21 \end{bmatrix}$$

$\vec{A} \quad \overset{\rightarrow}{x^*} \quad \vec{b}$

for least squares solution

$$A^T A \overset{\rightarrow}{x^*} = A^T \vec{b}$$

↳ this is least squares solution (closest fit for solution)

$$A^T A = \begin{bmatrix} -1 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 1 & 1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1+1+4 & -1+1+2 \\ -1+1+2 & 1+1+1 \end{bmatrix}$$

$$= \begin{bmatrix} 6 & 2 \\ 2 & 3 \end{bmatrix}$$

$$A^T b = \begin{bmatrix} -1 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 7 \\ 7 \\ 21 \end{bmatrix} = \begin{bmatrix} -7+7+42 \\ 7+7+21 \end{bmatrix} = \begin{bmatrix} 42 \\ 35 \end{bmatrix}$$

$$\begin{bmatrix} 6 & 2 \\ 2 & 3 \end{bmatrix} \overset{\rightarrow}{x^*} = \begin{bmatrix} 42 \\ 35 \end{bmatrix}$$

$$\Rightarrow \left[\begin{array}{cc|c} 6 & 2 & 42 \\ 2 & 3 & 35 \end{array} \right] \Rightarrow \begin{array}{l} 6m + 2c = 42 \\ 2m + 3c = 35 \end{array}$$

$$\begin{array}{r} \cancel{x_3} \quad \cancel{x_3} \quad \cancel{x_3} \\ - \quad - \quad - \\ -7c = -63 \end{array}$$

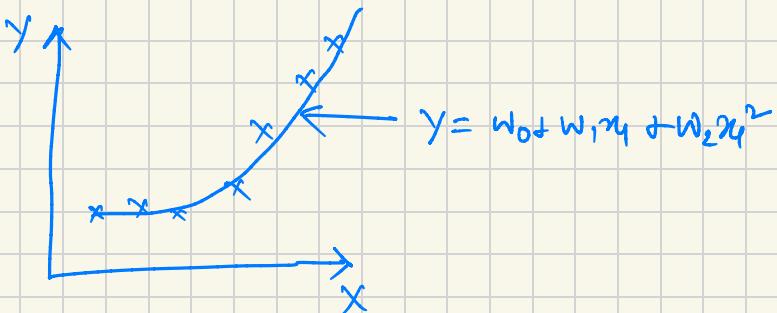
$$c = 9$$

$$m = 4$$

So, solution is $(9, 4)$. \checkmark

Polynomial regression:

It is a form of regression analysis in which the relationship between the independent variables and dependent variables are modeled in the nth degree polynomial.



Simple linear regression

$$y = w_0 + w_1x_1$$

multiple linear regression

$$y = w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n$$

polynomial linear regression

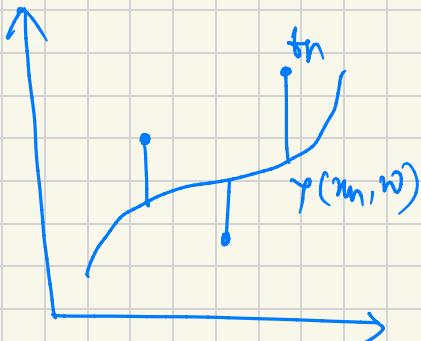
$$y = w_0 + w_1x_1 + w_2x_1^2 + \dots + w_nx_1^n$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{m+1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{m+1} \\ 1 & x_3 & x_3^2 & \cdots & x_3^{m+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{m+1} \end{bmatrix}_{n \times m} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix}_{m \times 1}$$

this can also be written as

$$\begin{bmatrix} \phi_0(x_1) \phi_1(x_1) \cdots \phi_{m-1}(x_1) \\ \phi_0(x_2) \phi_1(x_2) \cdots \phi_{m-1}(x_2) \\ \phi_0(x_3) \phi_1(x_3) \cdots \phi_{m-1}(x_3) \\ \vdots \\ \phi_0(x_n) \phi_1(x_n) \cdots \phi_{m-1}(x_n) \end{bmatrix}$$

so, $\boxed{y(x, w) = \sum_{i=0}^{m+1} w_i x^i}$



coefficients will be determined by fitting the polynomial to the training data. This can be done by minimizing an error function.

$$\begin{aligned} E(w) &= \frac{1}{2} \sum_{i=0}^N \{t_n - y(x, w)\}^2 \\ &= \frac{1}{2} \sum_{i=0}^N \{t_n - w^T \phi(x_i)\}^2 \end{aligned}$$

$$\nabla E(w) = \sum_{i=0}^N \{t_n - w^T \phi(x_i)\} \phi(x_i)^T \stackrel{?}{=} 0$$

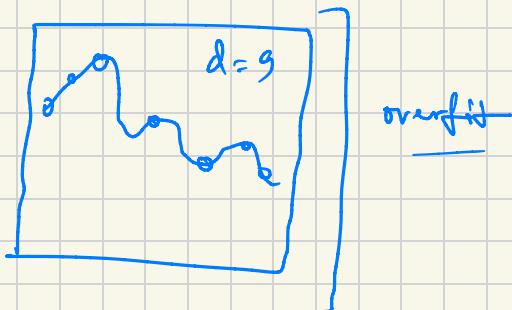
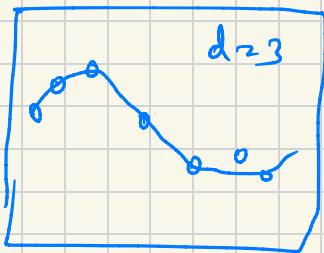
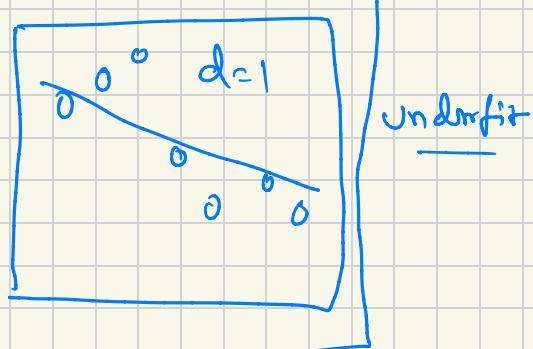
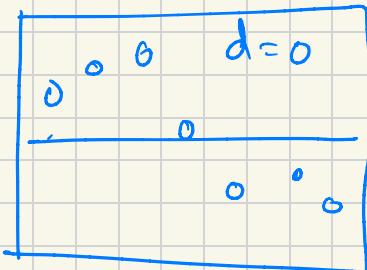
Convert to matrix form

$$(t - w^T \phi) \phi^T = 0$$

$$t \phi^T = w^T \phi \phi^T$$

$$\boxed{w = (\phi^T \phi)^{-1} \phi^T t}$$

~~*~~Choosing N~~~~: higher order polynomial provides excellent fit to training data but gives a very poor representation of the function.



Given the model complexity; the overfitting problem become less severe as the size of dataset increases.

$$E(w) = \frac{1}{2} \sum_{i=1}^N \left\{ t_i - w^T \phi(x_i) \right\}^2 + \frac{\lambda}{2} w^T w$$

This is loss term which is Least Square error. (LSE)

this penalty has been added to error function in order to discourage the coefficient from reaching large value.

~~* *~~ Regularization techniques are used to calibrate the linear regression models in order to minimize the loss function and prevent overfit & underfit.

If $\lambda = 0$, the whole model will overfit.

If $\lambda \gg 0$, the whole model will underfit.

More generalized form of regularizer is

$$E(w) = \frac{1}{N} \sum_{i=1}^N \left\{ t_i - w^T \phi(x_i) \right\}^2 + \frac{\lambda}{2} \sum_{j=1}^m |w_j|^q$$

If $q=2$ it is Rigid Regression

If $q=1$ it is Lasso Regression.

ElasticNet regularization [Both L1 & L2 regularization]

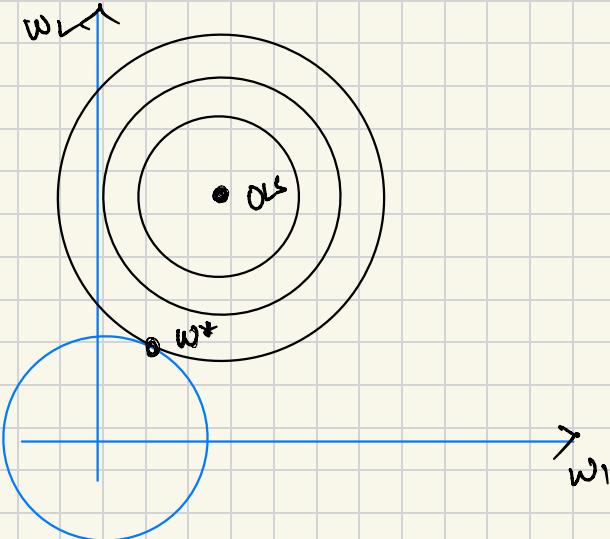
$$E(w) = \frac{1}{N} \sum_{i=1}^N \left\{ t_i - w^T \phi(x_i) \right\}^2 + \frac{\lambda}{2} \sum_{j=1}^m |w_j| + \frac{\lambda}{2} \sum_{j=1}^m |w_j|^2$$

$$\frac{\partial E(w)}{\partial w} = \sum_{i=1}^N \left\{ t_i - w^T \phi(x_i) \right\} \phi(x_i)^T + \frac{\lambda}{2} 2w^T = 0$$

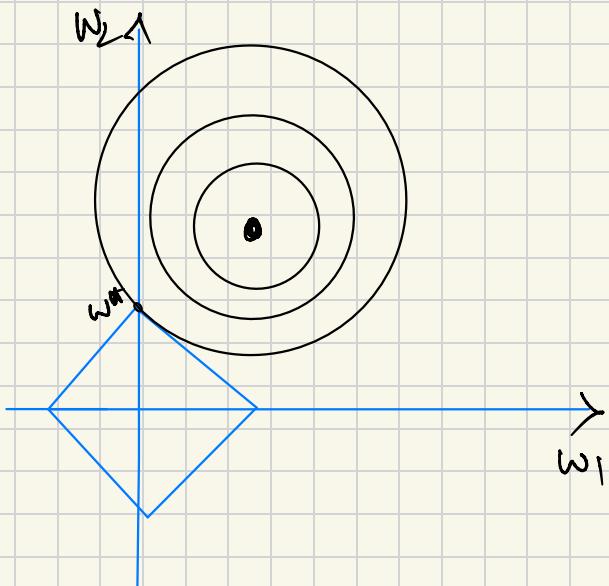
$$\Rightarrow t \phi(x)^T - w^T \phi(x) \phi(x)^T + \lambda w^T = 0$$

$$\Rightarrow w^T (\lambda I + \phi(x)^T \phi(x)) = \phi(x)^T t$$

$$\Rightarrow w = (\lambda I + \phi(x)^T \phi(x))^{-1} \phi(x)^T t$$



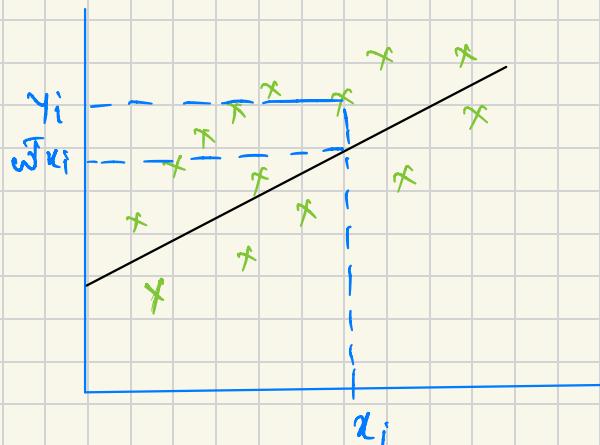
for $q=2$ (Ridge)



for $q=1$ (Lasso)

It gives sparse solution
for which $w_i^* = 0$

Probabilistic interpretation of Linear regression!



Likelihork

$$P(y_i | x_i) \sim N(w^T x_i, \sigma^2)$$

$$L(w) = \prod_{i=1}^n P(y_i | w^T x_i, \sigma^2)$$

MAP Estimation \rightarrow posterior \rightarrow likelihood \rightarrow prior

$$\boxed{P(w | z, y, \sigma^2, \beta^2)} \propto P(y | w^T x, \sigma^2) P(w | \beta^2)$$

$$\propto \prod_{i=1}^n P(y_i | w^T x_i, \sigma^2) \propto \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{w^T w}{2\sigma^2}}$$

let's think
it as a
gaussian
prior

$$\log L = \sum_{i=1}^n \log P(y_i | w^T x_i, \sigma^2) + \log \left(\frac{1}{\sqrt{2\pi}\beta} e^{-\frac{w^T w}{2\beta^2}} \right) N(w | 0, \beta^2)$$

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

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