1) Dual expression computation Ridge Regression

Soln

Using the slack variable $\xi_i = y_i - w^T x_i$

First we note that for minimization

$$\Rightarrow \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i$$ which is subject

to $y_i (w^T x_i) \geq \xi_i \quad \forall_i$

$\Rightarrow$ In this case we only have 2 opt variables w, ξ_i

$\Rightarrow$ The optimization problem then becomes,

1) Dual expression computation Ridge Regression

Soln

Using the slack variable $\xi_i = y_i - w^T x_i$

First we note that for minimization

$$\Rightarrow \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i$$ which is subject

to $y_i (w^T x_i) \geq \xi_i \quad \forall_i$

$\Rightarrow$ In this case we only have 2 opt variables w, ξ_i

$\Rightarrow$ The optimization problem then becomes,

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

Subject to

$$1 - \xi - y_i (w^T x_i) \leq 0 \quad \forall_i$$

Using the slack

First we note that for minimization

$\Rightarrow \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$ which is subject

to $y_i(w^T x_i) \geq \xi_i \quad \forall_i$

$\Rightarrow$ In this case we only have 2 opt variables

w, ξ_i

$\Rightarrow$ The optimization problem then becomes,

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

Subject to

$$1 - \xi - y_i(w^T x_i) \leq 0 \quad \forall_i$$

$$-\xi_i \leq 0 \quad \forall_i$$

The Langrangian of the above becomes

$$L(w, \xi, \mu, \lambda) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \mu_i(1-\xi$$

$\Rightarrow \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$

to $y_i(w^T x_i) \geq \xi_i \quad \forall_i$

$\Rightarrow$ In this case we only have 2 opt variables

w, ξ_i

$\Rightarrow$ The optimization problem then becomes,

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

Subject to

$$1 - \xi - y_i(w^T x_i) \leq 0 \quad \forall_i$$

$$-\xi_i \leq 0 \quad \forall_i$$

The Langrangian of the above becomes

$$L(w, \xi, \mu, \lambda) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \mu_i(1-\xi_i-y_i(x_i$$

$$- \sum_{i=1}^{n} \lambda_i \xi_i$$

We then take the derivative of L wrt ξ_i

to $y_i(w^T x_i)$

=) In this case we only w, ξ_i

=) The optimization problem then becomes,

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

Subject to

$$1 - \xi - y_i(w^T x_i) \le 0 \quad \forall_i$$

$$-\xi_i \le 0 \quad \forall_i$$

The Langrangian of the above becomes

$$L(w, \xi, \mu, \lambda) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \mu_i (1 - \xi_i - y_i(w^T x_i)) - \sum_{i=1}^{n} \lambda_i \xi_i$$

We then take the derivative of L w.r.t ξ_i to form the dual form, then

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

Subject to

$$1 - \xi - y_i(w^T x_i) \le 0 \quad \forall_i$$

$$-\xi_i \le 0 \quad \forall_i$$

The Langrangian of the above becomes

$$L(w, \xi, \mu, \lambda) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \mu_i (1 - \xi_i - y_i(w^T x_i)) - \sum_{i=1}^{n} \lambda_i \xi_i$$

We then take the derivative of L w.r.t ξ_i to form the dual form, then

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \mu_i + \lambda_i = C \quad \forall_i$$

=) $1 - \xi_i - y_i(w^T x_i) \le 0 \quad \forall_i$

This implies that

$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$

Subject to

$$1 - \xi_i - y_i(w^T x_i) \le 0 \quad \forall i$$
$$-\xi_i \le 0 \quad \forall i$$

The Lagrangian of the above becomes

$$L(w,\xi,\mu,\lambda) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\mu_i(1-\xi_i-y_i(w^T x_i)) - \sum_{i=1}^{n}\lambda_i\xi_i$$

We then take the derivative of L w.r.t ξ_i to form the dual form, then

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \mu_i + \lambda_i = C \quad \forall i$$

$$\Rightarrow 1 - \xi_i - y_i(w^T x_i) \le 0 \quad \forall i$$

$\Rightarrow$ this implies that

$$\mu_i(1 - \xi_i - y_i(w^T x_i)) = 0 \; ; \; \lambda_i \xi_i = 0 \quad \forall i$$

$\Rightarrow$ Thus for the complementary slackness condition we have

$$1 - y_i(w^T x_i) = 0$$

$\Rightarrow$ Finding the value for all μ_i the dual function takes the form

$$Q(\mu,\lambda) = \inf_{w,\xi} L(w,\xi,\mu,\lambda)$$

but from the Lagrangian we had that the term

$$\sum_i (C - \mu_i - \lambda_i)\xi_i$$

taking the infimum w.r.t ξ_i a

=> Thus for the complementary slackness condition we have

$$1 - y_i (w^T x_i) = 0$$

=> Finding the value for all μ_i the dual function takes the form

$$g(\mu, \lambda) = \inf_{w, \xi} L(w, \xi, \mu, \lambda)$$

but from the lagrangian we had that the term

$$\sum_i (C - \mu_i - \lambda_i) \xi_i$$

=> Finally taking the infimum w.r.t ξ_i and imposing $C - \mu_i - \lambda_i$ $\forall_i$ yields the dual problem

$$\max_{\lambda} q(\lambda) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j x_i^T x_j$$

with primal problem given by

takes the form

$$g(\mu, \lambda) = \inf_{w, \xi} L(w, \xi, \mu, \lambda)$$

but from the lagrangian we had that the term

$$\sum_i (C - \mu_i - \lambda_i) \xi_i$$

=> Finally taking the infimum w.r.t ξ_i and imposing $C - \mu_i - \lambda_i$ $\forall_i$ yields the dual problem

$$\max_{\lambda} q(\lambda) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j x_i^T x_j$$

with primal problem given by

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

Subject to

$$y_i (w^T x_i) \geq 1 - \xi_i \quad \text{and} \quad C = N \to \infty$$

but from the w, ξ Lagrangian we have that the term

$$\sum_i (C - \mu_i - \lambda_i)\,\xi_i$$

=) Finally taking the infimum w.r.t ξ_i and imposing $C - \mu_i - \lambda_i \quad \forall i$ yields the dual problem

$$\max_{\lambda} q(\lambda) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} \lambda_i \lambda_j y_i y_j x_i^T x_j$$

with primal problem given by

$$\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

Subject to

$$y_i (w^T x_i) \geq 1 - \xi_i \qquad \text{as } C = N \to \infty$$

) Dual ~~expression~~ C

2) Dual expression computation for SVM

soln

using the slack variable

$$\xi_i = y_i - w^T x_i - b$$

For the minimization $\frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$

Subject to

$$y_i (w^T x_i + b) \geq 1 - \xi_i \quad , \forall i$$

with $\xi_i \geq 0 \quad \forall i$

Opt variables in this case w, b, ξ_i

2) Dual expression computation for SVM

Using the slack variable

$$\xi_i = y_i - w^T x_i - b$$

For the minimization $\frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$

Subject to

$$y_i (w^T x_i + b) \geq 1 - \xi_i \quad , \forall_i$$

with $\xi_i \geq 0 \quad \forall_i$

Opt variables in the case w, b, ξ_i

⇒ Note that for all poible deviations for change in $w^T x$ and b we have.

$$w^T x + b = 1$$
$$w^T x + b = 0$$
$$w^T x + b = -1$$

⇒ We then consider $w^T x + b = 0$, and let the

Subject to

$$y_i (w^T x_i + b) \geq 1 - \xi_i \quad , \forall_i$$

with $\xi_i \geq 0 \quad \forall_i$

Opt variables in the case w, b, ξ_i

⇒ Note that for all poible deviations for change in $w^T x$ and b we have.

$$w^T x + b = 1$$
$$w^T x + b = 0$$
$$w^T x + b = -1$$

⇒ We then consider $w^T x + b = 0$, and let the

optimization problem be given as

$$\min_{w, b, \xi_i} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$$

Note $w^T w =$

Subject to

⇒ Note that for all possible deviations for change in $w^T x$ and b we have.

$$w^T x + b = 1$$
$$w^T x + b = 0$$
$$w^T x + b = -1$$

⇒ We then consider $w^T x + b = 0$, and let the optimization problem be given as

$$\min_{w, b, \xi_i} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i \qquad \text{note } w^T w = \|w\|^2$$

Subject to

$$1 - \xi_i - y_i (w^T x_i + b) \leq 0 \quad \forall_i$$
$$-\xi_i \leq 0 \quad \forall_i$$

⇒ Thus the Lagrangian becomes

$$L(w, b, \xi, \mu, \lambda) = \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i + \cdots +$$

$$+ \sum_{i=1}^{n} \mu_i (1 - \xi_i - y_i (w^T x_i + b)) - \sum_{i=1}^{n} \lambda_i \xi_i$$

$\Rightarrow$ Taking derivatives of L w.r.t b and ξ respectively we have

$$\nabla_w L = 0 \Rightarrow w^* = \sum_{i=1}^{n} \mu_i x y_i x_i$$

$$\Rightarrow \frac{\partial L}{\partial b} = 0 \Rightarrow \sum \mu_i x y_i = 0$$

$$\Rightarrow \frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \mu_i x \lambda_i = C \quad \forall_i$$

$$+ \sum_{i=1}^{n} \mu_i (1 - \xi_i - y_i (w^T x_i + b)) - \sum_{i=1}^{n} \lambda_i \xi_i$$

$\Rightarrow$ Taking derivatives of L w.r.t b and ξ respectively we have

$$\nabla_w L = 0 \Rightarrow w^* = \sum_{i=1}^{n} \mu_i x y_i x_i$$

$$\Rightarrow \frac{\partial L}{\partial b} = 0 \Rightarrow \sum \mu_i x y_i = 0$$

$$\Rightarrow \frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \mu_i x \lambda_i = C \quad \forall_i$$

$$\Rightarrow 1 - \xi_i - y_i (w^T x_i + b) \leq 0 \; ; \; \xi_i \geq 0 \quad \forall_i$$

$$\Rightarrow \mu_i (1 - \xi_i - y_i (w^T x_i + b) = 0 \; ; \; \lambda_i \xi_i = 0 \quad \forall_i$$

$\Rightarrow$ For the complementary slackness condition, we h

we have

$\nabla_w L = 0 \Rightarrow w^* = \sum_{i=1}^{n} \mu_i x y_i x_i$

$\Rightarrow \frac{\partial L}{\partial b} = 0 \Rightarrow \sum \mu_i x y_i = 0$

$\Rightarrow \partial L / \partial \xi_i = 0 \Rightarrow \mu_i x \lambda_i = C \quad \forall_i$

$\Rightarrow 1 - \xi_i - y_i (w^T x_i + b) \leq 0 \; ; \; \xi_i \geq 0 \quad \forall_i$

$\Rightarrow \mu_i (1 - \xi_i - y_i (w^T x_i + b) = 0 \; ; \; \lambda_i \xi_i = 0 , \forall_i$

$\Rightarrow$ For the complementary slackness condition, we have that

$1 - y_i (w^T x_i + b) = 0 \quad —①$

$b = y_i - x_j^T w$, j such that $0 < \mu_j < C \quad —②$

$\Rightarrow$ Thus finding value for all μ_i, the dual functi

$\Rightarrow \frac{\partial L}{\partial b} = 0 \Rightarrow \sum \mu_i x y_i = 0$

$\Rightarrow \partial L / \partial \xi_i = 0 \Rightarrow \mu_i x \lambda_i = C \quad \forall_i$

$\Rightarrow 1 - \xi_i - y_i (w^T x_i + b) \leq 0 \; ; \; \xi_i \geq 0 \quad \forall_i$

$\Rightarrow \mu_i (1 - \xi_i - y_i (w^T x_i + b) = 0 \; ; \; \lambda_i \xi_i = 0 , \forall_i$

$\Rightarrow$ For the complementary slackness condition, we have that

$1 - y_i (w^T x_i + b) = 0 \quad —①$

$b = y_i - x_j^T w$, j such that $0 < \mu_j < C \quad —②$

$\Rightarrow$ Thus finding value for all μ_i, the dual function becomes

$$Q(\mu, \lambda) = \inf_{w, b, \xi} L(w, b, \xi, \mu, \lambda)$$

$\Rightarrow 1 - \xi_i - y_i(w^T x_i + b) \leq 0 \quad ; \quad \xi_i \geq 0 \quad \forall_i$

$\Rightarrow \mu_i(1 - \xi_i - y_i(w^T x_i + b)) = 0 \quad ; \quad \lambda_i \xi_i = 0 \quad , \forall_i$

$\Rightarrow$ For the complementary slackness condition, we have that

$$1 - y_i(w^T x_i + b) = 0 \quad \text{---(1)}$$

$$b = y_i - x_j^T w \quad , j \text{ such that } 0 < \mu_j < C \quad \text{---(2)}$$

$\Rightarrow$ Thus finding value for all μ_i, the dual function becomes

$$q(\mu, \lambda) = \inf_{w, b, \xi} L(w, b, \xi, \mu, \lambda)$$

but from lagrangian we have the term

$$\sum_i (C - \mu_i - \lambda_i)\xi_i$$

$\Rightarrow$ Now taking the infimum w.r.t ξ_i and imposing $C = \mu_i + \lambda_i \quad \forall_i$, the dual problem becomes

$$\max_{\mu} q(\mu) = \sum_{i=1}^{n} \mu_i - \frac{1}{2} \sum_{i,j=1}^{n} \mu_i \mu_j y_i y_j x_j^T x_j$$

Subject to

$$0 \leq \mu_i \leq C \quad i = 1, \ldots n \qquad \sum_{i=1}^{n} y_i \mu_i = 0$$

And for the primal problem we have

$$\text{minimize } \frac{1}{2} w^T w + C \sum_i \xi_i$$

=) Now taking the infimum w.r.t ξ_i and imposing $C = \alpha_i + \lambda_i$ $\forall_i$, the dual problem becomes;

$$\max_{\alpha} \; q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_j^T x_j$$

Subject to

$$0 \le \alpha_i \le C \quad i = 1, \dots n \qquad \sum_{i=1}^{n} y_i \alpha_i = 0$$

and for the primal problem we have

$$\text{Minimize}_p \; \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$$

Subject to

$$y_i (w^T x_i + b) \ge 1 - \xi_i \quad \forall_i$$

as $\epsilon = N \to \infty$

How to obtain

$$\max_{\alpha} \; q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_j^T x_j$$

Subject to

$$0 \le \alpha_i \le C \quad i = 1, \dots n \qquad \sum_{i=1}^{n} y_i \alpha_i = 0$$

and for the primal problem we have

$$\text{Minimize}_p \; \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$$

Subject to

$$y_i (w^T x_i + b) \ge 1 - \xi_i \quad \forall_i$$

as $\epsilon = N \to \infty$

(c) How to obtain prediction for ~~x~~

Soln

=) We let K be a $n \times n$ matrix with $K_{ij} = k(x_i, x_j)$ such that K is positive semi-definite $\forall n$ and all data sets $\{x_1, \dots x_n\}$

and for the primal problem we have

minimize $\frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$

Subject to

$y_i (w^T x_i + b) \geq 1 - \xi_i \quad \forall_i$

as $\epsilon = N \to \infty$

3 (a) How to obtain prediction for x_t

Soln

⇒ We let k be a $n \times n$ matrix with $k_{ij} = k(x_i, x_j)$ such that k is positive semi-definite $\forall n$ and all data sets $\{x_1, \ldots x_n\}$

⇒ Thus, given any n, and feature vectors $x_1, \ldots x_n$ we will have that for all scalars $c_1, \ldots c_n$

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0 \quad \longleftarrow (1)$$

minimize $\frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$

Subject to

$y_i (w^T x_i + b) \geq 1 - \xi_i \quad \forall_i$

as $\epsilon = N \to \infty$

3 (a) How to obtain prediction for x_t

Soln

⇒ We let k be a $n \times n$ matrix with $k_{ij} = k(x_i, x_j)$ such that k is positive semi-definite $\forall n$ and all data sets $\{x_1, \ldots x_n\}$

⇒ Thus, given any n, and feature vectors $x_1, \ldots x_n$ we will have that for all scalars $c_1, \ldots c_n$

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0 \quad \longleftarrow (1)$$

⇒ Now for the kernel function given below

$$k(x_i, x_j) = \exp\left(-\|x_i - x_j\|^2\right) / 2\sigma^2$$

Subject to $y_i (w^T x_i + b) \geq 1 - \xi_i$ $\forall_i$ as $C = N \to \infty$

3 (a) How to obtain prediction for x_t

Soln

⇒ We let K be a $n \times n$ matrix with $k_{ij} = k(x_i, x_j)$ such that K is positive semi-definite $\forall n$ and all data sets $\{x_1 \dots x_n\}$

⇒ Thus, given any n, and feature vectors $x_1 \dots x_n$ we will have that for all scalars $c_1 \dots c_n$

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) = 0 \quad \longleftarrow (1)$$

⇒ Now for the kernel function given below

$$k(x_i, x_j) = \exp\left(-\|x_i - x_j\|^2\right) / 2\sigma^2$$

any data values x_i and x_j feed forward into the function (1) will yield the result of the predicted x_t values

⇒ This can be generally classified by the ~~function~~ function $f(x)$ for x_t prediction as

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i k(x_i, x_j) + b$$

$$= \sum_{i=1}^{n} \alpha_i y_i e^{\frac{\|x_i - x_j\|^2}{2\sigma^2}} + b$$

=) This can be generally classified by the ~~function~~ function $f(x)$ for x_t prediction as

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i k(x_i, x_j) + b$$

$$= \sum_{i=1}^{n} \alpha_i y_i e^{\frac{\|x_i - x_j\|^2}{2\sigma^2}} + b$$

4) Maximum likelihood Normal distribution

=) Compute MAP to the MLE estimation (MAP MLE) as $N \to \infty$

Soln

Given $x_1, x_2, x_3, \ldots x_n$ for estimation we know that the p.d.f for normal distribution is generally

$$= \sum_{i=1}^{n} \alpha_i y_i e^{\frac{\|x_i - x_j\|^2}{2\sigma^2}} + b$$

4) Maximum likelihood Normal distribution

=) Compute MAP to the MLE estimation (MAP MLE) as $N \to \infty$

Soln

Given $x_1, x_2, x_3, \ldots x_n$ for estimation we know that the p.d.f for normal distribution is generally given by

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

) If we apply the equation to n - number of observation we have.

$$\frac{1}{(x_i - \mu)^2}$$

4) Maximum likelihood Normal distribution
=) Compute MAP to the MLE estimation
(MAPLE) as $N \to \infty$

Soln

Given $x_1, x_2, x_3, \ldots x_n$ for estimation we know that the p.d.f for normal distribution is generally given by

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

=) If we apply the equation to n-number of observation we have

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

$$L(\mu, x_i, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^3}{2\sigma^2}}$$

=) Compute MAP to the MLE estimation
(MAPLE) as $N \to \infty$

Soln

Given $x_1, x_2, x_3, \ldots x_n$ for estimation we know that the p.d.f for normal distribution is generally given by

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

=) If we apply the equation to n-number of observation we have

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

$$L(\mu, x_i, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^3}{2\sigma^2}}$$

=) Simplifying this we obtain

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot \cdots \cdot \frac{1}{\sigma\sqrt{2\pi}} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n$$

$$= e^{-\frac{1}{2\sigma^2}(x_1-\mu)^2} \cdot e^{-\frac{1}{2\sigma^2}(x_2-\mu)^2} \cdot \cdots \cdot e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

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

Taking the log of the above function we have

=) Simplifying this we obtain

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot \cdots \cdot \frac{1}{\sigma\sqrt{2\pi}} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n$$

$$= e^{-\frac{1}{2\sigma^2}(x_1-\mu)^2} \cdot e^{-\frac{1}{2\sigma^2}(x_2-\mu)^2} \cdot \cdots \cdot e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

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

=) Taking the log of the above function we have

$$\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n + \log\left(e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2}\right)$$

$$\Rightarrow \frac{1}{\sigma\sqrt{2\pi}} \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot \cdots \cdot \frac{1}{\sigma\sqrt{2\pi}} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n$$

$$= e^{-\frac{1}{2\sigma^2}(x_1-\mu)^2} \cdot e^{-\frac{1}{2\sigma^2}(x_2-\mu)^2} \cdot \cdots \cdot e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

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

$\Rightarrow$ Taking the log of the above function we have

$$\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n + \log\left(e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2}\right)$$

$$\Rightarrow n\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + -\frac{1}{2\sigma^2}\sum(x_i-\mu)^2 = n\log(1) - n\log(\sigma(2\pi$$

$$- \frac{1}{2\sigma^2}\sum(x_i-\mu)^2$$

$$= e^{-\frac{1}{2\sigma^2}(x_1-\mu)^2} \cdot e^{-\frac{1}{2\sigma^2}(x_2-\mu)^2} \cdot \cdots \cdot e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

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

$\Rightarrow$ Taking the log of the above function we have

$$\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n + \log\left(e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2}\right)$$

$$\Rightarrow n\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + -\frac{1}{2\sigma^2}\sum(x_i-\mu)^2 = n\log(1) - n\log(\sigma(2\pi)) -$$

$$- \frac{1}{2\sigma^2}\sum(x_i-\mu)^2$$

$$= -n\log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2$$

$$= \frac{2L}{} = \frac{2L}{} = -\frac{1}{} \sum^{n} (x_i-\mu) \cdot 2 \cdot -1$$

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

$\Rightarrow$ Taking the log of the above function we have

$$\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n + \log\left(e^{-\frac{1}{2\sigma^2}\sum (x_i-\mu)^2}\right)$$

$$\Rightarrow n\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \frac{-1}{2\sigma^2}\sum(x_i-\mu)^2 = n\log(1) - n\log(\sigma(2\pi)) - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2$$

$$= -n\log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2$$

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

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

$\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n + \log\left(e^{-\frac{1}{2\sigma^2}\sum (x_i-\mu)^2}\right)$

$\Rightarrow n\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \frac{-1}{2\sigma^2}\sum(x_i-\mu)^2 = n\log(1) - n\log(\sigma(2\pi)) - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2$

$= -n\log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2$

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

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

$$\Rightarrow \frac{1}{\sigma^2}\sum x_i - \frac{1}{\sigma^2}\mu = 0 \quad = \frac{1}{\sigma^2}\mu = \frac{1}{\sigma^2}\sum x_i$$

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

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

5) Decision Trees

a) Discussion.

In this problem we gonna be discussing our decision tree information gain based on Entropy which is used to determine or Calculate the purity of a substance. In other words we going to use Information gain to help in getting and understanding its role in building decision tree by getting the total average for entropy based on each split in the below example

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

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

5) Decision Trees

a) Discussion.

In this problem we gonna be discussing our decision tree information gain based on Entropy which is used to determine or Calculate the purity of a substance. In other words we going to use Information gain to help in getting and understanding its role in building decision tree by getting the total average for entropy based on each split in the below example

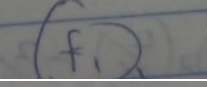
5) Decision Trees

a) Discussion

In this problem we gonna be discussing our decision tree information gain based on Entropy which is used to determine or Calculate the purity of a substance

=> In other words we going to use Information gain to help in getting and understanding its role in building decision, tree by getting the total average for entropy based on each specific split in the below example

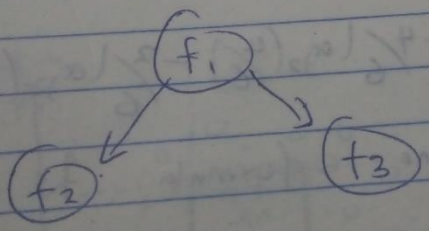
e.g Lets say we are splitting a function (Fi) as shown below.



In this problem we gonna be discussing our decision tree information gain based on Entropy which is used to determine or Calculate the purity of a substance

=> In other words we going to use Information gain to help in getting and understanding its role in building decision, tree by getting the total average for entropy based on each specific split in the below example

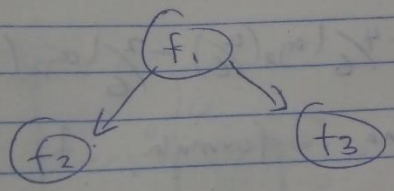
e.g Lets say we are splitting a function (Fi) as shown below.



that Gain function is given by

=> In other words we going to use Information gain to help in getting and understanding its role in building decision tree by getting the total average for entropy based on each specific split in the below example

e.g Let's say we are splitting a function (F_i) as shown below -

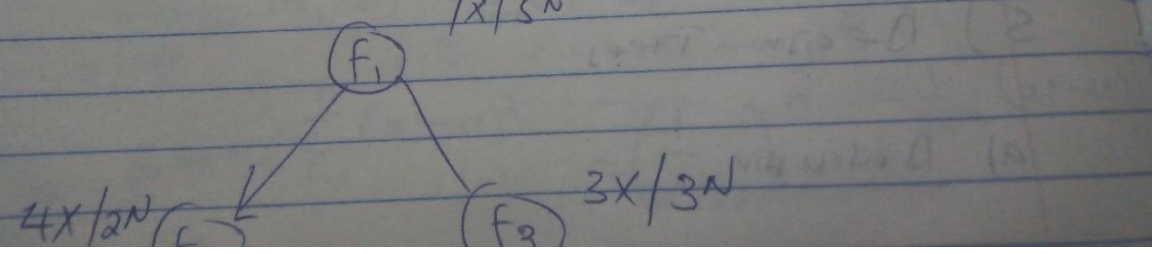


Note that Gain function is given by

$$Gain(S, A) = H(S) - \sum_{v \in Val} \frac{S_v}{S} H(S_v)$$

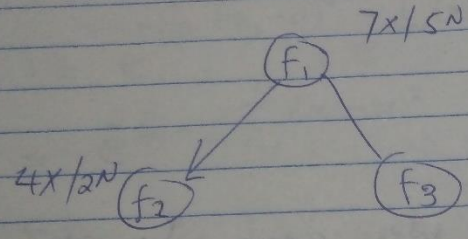
where S_v is the subset after splitting and
S is the total sub set

=> Now, Lets say we have to split (f_1) and we are splitting it based on (f_2) and (f_3) as shown below



where S_v is the subset after splitting and
S is the total subset

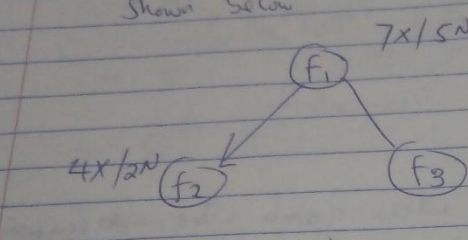
=) Now, lets say we have to split f_1 and we are splitting it based on f_2 and f_3 as shown below



Assume also we have a 7X/5N
=) 7-Values and 5N obs; total
and distributed as shown above

=) Now, the first step is to compute H(S) which is basically for the Subset f_1 = H(f_1) ≃ H(S)

=) Now, lets say we have to split f_1 and we are splitting it based on f_2 and f_3 as shown below



Assume also we have a 7X/5N
=) 7-Values and 5N obs; total
and distributed as shown above

=) Now, the first step is to compute H(S) which is basically for the Subset f_1 = H(f_1) ≃ H(S)
which is the root node

=) Then applying the formula

$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

4X/2N (f2) (f3) 3X/3N

Assume also we have 7X/5N
=) 7 - Values and 5N nodes that
and distributed as shown above.

=) Now the first step is to compute H(S) which
is basically for the Subset (f1) = H(f1) ≃ H(S)
which is the root node

=) Then applying the formula

$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

$$= -4/6 \log_2(4/6) - 3/6 \log_2(3/6) \simeq 0.87$$

=) We use the same formula to obtain H(f2) = 0?
and H(f3) = 1
=) This is because we have values ranging
from 0 ⟶ 1

Assume also we have 7X/5N
and distributed as shown above.

=) Now the first step is to compute H(S) which
is basically for the Subset (f1) = H(f1) ≃ H(S)
which is the root node

=) Then applying the formula

$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

$$= -4/6 \log_2(4/6) - 3/6 \log_2(3/6) \simeq 0.87$$

=) We use the same formula to obtain H(f2) = 0.?
and H(f3) = 1
=) This is because we have values ranging
from 0 ⟶ 1
=) 1 => for completly impurity
0 -> For complete purity

=) Since we are computing H(S) from root node, the we have

$$Gain(S, f_1) = H(S) - \frac{6}{12} H(f_2) - \frac{6}{12} H(f_3)$$

$$= 0.87 - \frac{6}{12} \times 0.73 - \frac{6}{12} \times 1 = 0.005$$

=) So basically if the above split gives the highest information gain value then the above structure will actually be used to split the particular dataset itself inorder to construct the decision tree

=) The higher the information gain the higher chance that spliting criterion will be used for constructing the decision tree

=) So incase we use conditional gain, then minimizing the conditional gain makes sense to build a decision tree.

=) Since we are computing H(S) from root node, the we have

$$Gain(S, f_1) = H(S) - \frac{6}{12} H(f_2) - \frac{6}{12} H(f_3)$$

$$= 0.87 - \frac{6}{12} \times 0.73 - \frac{6}{12} \times 1 = 0.005$$

=) So basically if the above split gives the highest information gain value then the above structure will actually be used to split the particular dataset itself inorder to construct the decision tree

=) The higher the information gain the higher chance that spliting criterion will be used for constructing the decision tree

=) So incase we use conditional gain, then minimizing the conditional gain makes sense to build a decision tree.

$$Gain(S, f_1) = H(S) - \frac{6}{12} H(f_2) - \frac{6}{12} H(f_3)$$

$$= 0.87 - \frac{6}{12} \times 0.73 - \frac{6}{12} \times 1 = 0.005$$

=> So basically if the above split gives the highest information gain value then the above structure will actually be used to split the particular dataset itself inorder to construct the decision tree.

=> The higher the information gain the higher chance that splitting criterion will be used for constructing the decision tree.

=> So incase we use conditional gain, then minimizing the conditional gain makes sense to greedily build a decision tree.

b) With the given data points to draw the decision boundary we have that

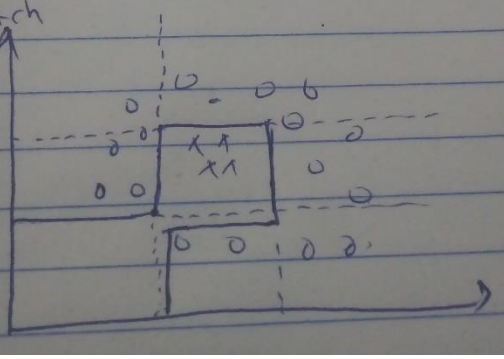
particular dataset itself inorder to construct the decision tree.

=> The higher the information gain the higher chance that splitting criterion will be used for constructing the decision tree.

=> So incase we use conditional gain, then minimizing the conditional gain makes sense to greedily build a decision tree.

b) With the given data points to draw the decision boundary we have that

a sketch



- Understanding how decision tree works is that decision boundaries will always be perpendicular to X and Y axis as for that case we made the above estimation

⇒ So, basically the above sketch of the model can be said to be overfitted and this happens when we increase the depth of the tree
⇒ This tells us that the model above is prone to overfitting

⇒ Thus, we see that if no limit on the depth of the algorithm, then this will be prone to overfitting implying that we going to have a complex decision boundary accomodating

- Understanding how decision tree works is that decision boundaries will always be perpendicular to X and Y axis as for that case we made the above estimation

⇒ So, basically the above sketch of the model can be said to be overfitted and this happens when we increase the depth of the tree
⇒ This tells us that the model above is prone to overfitting

⇒ Thus, we see that if no limit on the depth of the algorithm, then this will be prone to overfitting implying that we going to have a complex decision boundary accomodating

=> Understanding how decision tree works is that decision boundaries will always be perpendicular to X and Y axis as for that case we made the above estimation.

=> So, basically the above sketch of the model can be said to be overfitted when we increase the depth of the tree and this happens
=> This tells us that the model above is prone to overfitting

=> Thus, we see that if no limit on the depth of the algorithm, then this will be prone to overfitting implying that we going to have a complex decision boundary accomodating all similar data points

(C) (C) Increase the value of max depth may overfit the data.

made the above estimation

=> So, basically the above sketch of the model can be said to be overfitted when we increase the depth of the tree and this happens
=> This tells us that the model above is prone to overfitting

=> Thus, we see that if no limit on the depth of the algorithm, then this will be prone to overfitting implying that we going to have a complex decision boundary accomodating all similar data points

(C) (C) Increase the value of max depth may overfit the data.

END