Discussion Section - Week 2

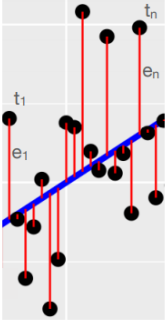
10/14/2020

Xuehai He

What we have learnt this week?

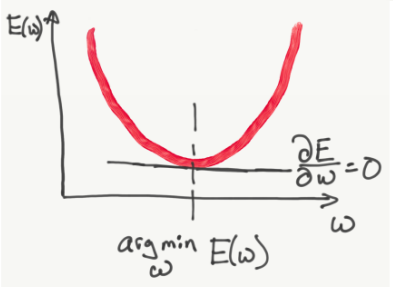
- Linear Regression with Ordinary Least Squares Error
- Data splitting: cross-validation setup
- Decision boundaries

Linear Regression & Ordinary Least Squares Error



$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

$$E(\omega) = \frac{1}{2}\left(X\omega - t\right)^2$$

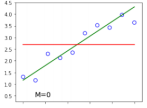


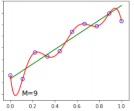
Solving for w by differentiating E(w) w.r.t. w and equating it to zero we get,

$$\omega = (X^T X)^{-1} X^T t$$

Data splitting: cross-validation setup

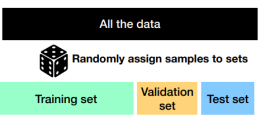
- Bias – measures the difference between predicted & expected value
- Variance – measures how well the model can deal with fluctuations in the data.





Green line – data generating fn
Red line – fitted fn

k-fold splitting



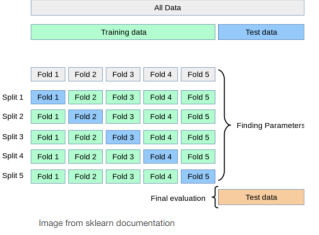


Image from sklearn documentation

As k increases, variance increases

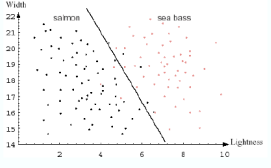
Decision boundaries - Classifiers

$$S_{training} = \{(\mathbf{x}_i, y_i), i = 1..n\} \qquad \mathbf{x} = (x_1, ..., x_m), x_i \in \mathbb{R}, \quad \mathbf{x} \in \mathbb{R}^m$$

$$y \in \{0, 1\} \qquad y = 0\text{: negative} \qquad y = 1\text{: positive}$$

Classifier: $f(\mathbf{x}; W) \in \{0, 1\}$

Model parameter to be learned: W



Training error:

$$e_{training} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(y_i \neq f(\mathbf{x}_i; W)) \qquad \text{(0-1 loss)}$$

Step 1 - Training

- We train a supervised classifier that learns by minimizing misclassification error (cost/loss).
- The error is normally defined as

$$\text{Minimize } \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(y_i \neq f(\mathbf{x}_i; W))$$

Step 2 - Validating

- We test the performance on the validation set to pick the best model that does not overfit to the training data

Step 3 - Testing

- **The difference between the testing and training set error is called generalization error.**
- To minimize the error, we should increase the training set size to help the model to generalize.
- Exploit the cross-validation setup!

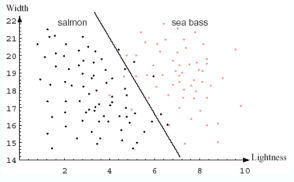
For decision boundary

Let **x** be the input vector (observation) and y be its label:

Often, we are given a set of training data

$$S = \{(\mathbf{x}_i, y_i), i = 1..n\} \quad \mathbf{x} = (x_1, ..., x_m), x_i \in \mathcal{R}, \quad \mathbf{x} \in \mathcal{R}^m$$

A classifier f(**x**):



Decision boundary:
$\{\mathbf{x}_i, f(\mathbf{x}_i) = 0\}$

For decision boundary

- **The decision boundary of a binary classifier refers to the set of data samples that are "on the fence" between making the decision being positive or negative: that is being 50%-50% for classification.**

- **For a linear model, the decision boundary is a line/hyper-plane, depending upon the dimension of the data.**

- **There is often a bias terms, b (scalar), refers to as the translation (shift) of the decision boundary.**

What we are learning today?

Still Mathematical things

Just to strength your skills in Vector Calculus

So Once Again

Once Again

Derivatives with vectors (Numerator layout)

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix}^\mathsf{T} \qquad \mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^\mathsf{T}$$

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix} \qquad \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

“Jacobian formulation”

https://en.wikipedia.org/wiki/Matrix_calculus

Once Again

Derivatives with vectors (Denominator layout)

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix}^\mathsf{T} \qquad \mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^\mathsf{T}$$

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix}$$

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

“Hessian formulation”

https://en.wikipedia.org/wiki/Matrix_calculus

Vector-by-scalar

$$\mathbf{y}(x) = \begin{pmatrix} y_1(x) & y_2(x) & y_3(x) \end{pmatrix}$$

$$\frac{d\mathbf{y}(x)}{dx} = \begin{pmatrix} \frac{dy_1(x)}{dx} & \frac{dy_2(x)}{dx} & \frac{dy_3(x)}{dx} \end{pmatrix}$$

Vector-by-vector

$$\mathbf{y}(\mathbf{x}) = \begin{pmatrix} y_1(\mathbf{x}) & ,..., & y_m(\mathbf{x}) \end{pmatrix}$$

$$\mathbf{x} = \begin{pmatrix} x_1 & ,..., & x_n \end{pmatrix}$$

$$\frac{d\mathbf{y}(\mathbf{x})}{d\mathbf{x}} = \begin{pmatrix} \frac{dy_1(\mathbf{x})}{dx_1} & ,..., & \frac{dy_m(\mathbf{x})}{dx_1} \\ . & . & . \\ \frac{dy_1(\mathbf{x})}{dx_n} & ,..., & \frac{dy_m(\mathbf{x})}{dx_n} \end{pmatrix}$$

$$A = \begin{pmatrix} a_{11} & ,..., & a_{1m} \\ . & . & . \\ a_{n1} & ,..., & a_{nm} \end{pmatrix}$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ . \\ x_m \end{pmatrix}$$

Vector-by-vector

$$\frac{\partial A\mathbf{x}}{\partial \mathbf{x}} = A$$

numerator form

$$\frac{\partial \mathbf{x}^T A^T}{\partial \mathbf{x}} = A$$

denominator form

Identities: vector-by-vector $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$

Condition	Expression	Numerator layout, i.e. by y and $\mathbf{x}^\mathsf{T}$	Denominator layout, i.e. by $\mathbf{y}^\mathsf{T}$ and x
a is not a function of **x**	$\frac{\partial \mathbf{a}}{\partial \mathbf{x}} =$	$\mathbf{0}$	
	$\frac{\partial \mathbf{x}}{\partial \mathbf{x}} =$	$\mathbf{I}$	
A is not a function of **x**	$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} =$	$\mathbf{A}$	$\mathbf{A}^\top$
A is not a function of **x**	$\frac{\partial \mathbf{x}^\top \mathbf{A}}{\partial \mathbf{x}} =$	$\mathbf{A}^\top$	$\mathbf{A}$
a is not a function of **x**, **u** = **u**(**x**)	$\frac{\partial a\mathbf{u}}{\partial \mathbf{x}} =$	$a\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	
a = *a*(**x**), **u** = **u**(**x**)	$\frac{\partial a\mathbf{u}}{\partial \mathbf{x}} =$	$a\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \mathbf{u}\frac{\partial a}{\partial \mathbf{x}}$	$a\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial a}{\partial \mathbf{x}}\mathbf{u}^\top$
A is not a function of **x**, **u** = **u**(**x**)	$\frac{\partial \mathbf{A}\mathbf{u}}{\partial \mathbf{x}} =$	$\mathbf{A}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\mathbf{A}^\top$
u = **u**(**x**), **v** = **v**(**x**)	$\frac{\partial (\mathbf{u}+\mathbf{v})}{\partial \mathbf{x}} =$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$	
u = **u**(**x**)	$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{x}} =$	$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}$

Matrix-by-scalar

$$Y(x) = \begin{pmatrix} y_{11}(x) & , ..., & y_{1m}(x) \\ . & . & . \\ y_{n1}(x) & , ..., & y_{nm}(x) \end{pmatrix}$$

$$\frac{dY(x)}{dx} = \begin{pmatrix} \frac{dy_{11}(x)}{dx} & , ..., & \frac{dy_{1m}(x)}{dx} \\ \frac{dy_{n1}(x)}{dx} & , ..., & \frac{dy_{nm}(x)}{dx} \end{pmatrix}$$

Scalar-by-vector

$$A = \begin{pmatrix} a_{11} & ,..., & a_{1n} \\ . & . & . \\ a_{n1} & ,..., & a_{nn} \end{pmatrix}$$

$$\mathbf{x} = \begin{pmatrix} x_1 \\ . \\ x_n \end{pmatrix}$$

$$\frac{\partial \mathbf{x}^T A \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{x}^T A$$

Matrix calculus

Condition	Expression	Numerator layout, i.e. by $\mathbf{x}^\mathsf{T}$; result is row vector	Denominator layout, i.e. by x; result is column vector
a is not a function of **x**	$\frac{\partial(\mathbf{a}\cdot\mathbf{x})}{\partial\mathbf{x}} = \frac{\partial(\mathbf{x}\cdot\mathbf{a})}{\partial\mathbf{x}} = \frac{\partial\mathbf{a}^\mathsf{T}\mathbf{x}}{\partial\mathbf{x}} = \frac{\partial\mathbf{x}^\mathsf{T}\mathbf{a}}{\partial\mathbf{x}} =$	$\mathbf{a}^\mathsf{T}$	$\mathbf{a}$
A is not a function of **x** **b** is not a function of **x**	$\frac{\partial\mathbf{b}^\mathsf{T}\mathbf{A}\mathbf{x}}{\partial\mathbf{x}} =$	$\mathbf{b}^\mathsf{T}\mathbf{A}$	$\mathbf{A}^\mathsf{T}\mathbf{b}$
A is not a function of **x**	$\frac{\partial\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}}{\partial\mathbf{x}} =$	$\mathbf{x}^\mathsf{T}(\mathbf{A}+\mathbf{A}^\mathsf{T})$	$(\mathbf{A}+\mathbf{A}^\mathsf{T})\mathbf{x}$
A is not a function of **x** **A** is symmetric	$\frac{\partial\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}}{\partial\mathbf{x}} =$	$2\mathbf{x}^\mathsf{T}\mathbf{A}$	$2\mathbf{A}\mathbf{x}$
A is not a function of **x**	$\frac{\partial^2\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}}{\partial\mathbf{x}^2} =$	$\mathbf{A}+\mathbf{A}^\mathsf{T}$	
A is not a function of **x** **A** is symmetric	$\frac{\partial^2\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}}{\partial\mathbf{x}^2} =$	$2\mathbf{A}$	

Problem: find $\frac{\partial E(w)}{\partial w}$ for

$$E(w) = \frac{1}{2}(X\vec{w} - \vec{t})^2$$

→ OLS error

X = $N \times M$ matrix

$\vec{w}$ = $M \times 1$ vector

$\vec{t}$ = $M \times 1$ vector

{ N datapoints
M features }

$$E(w) = \frac{1}{2}(Xw - t)^2$$

$$= \frac{1}{2}(Xw - t)^T (Xw - t)$$

$$= \frac{1}{2}(w^T X^T - t^T)(Xw - t)$$

$$= \frac{1}{2}(w^T X^T X w - t^T X w - w^T X^T t - t^T t$$

$$= \frac{1}{2}(w^T X^T X w - 2 w^T X^T t)$$

$$\frac{\partial E(w)}{\partial w} = \frac{1}{2}(2 X^T X w - 2 X^T t)$$

QA time

10/14/2020