6. Statistics and Linear Algebra Preliminaries — Learning Apache Spark with Python documentation

Learning Apache Spark with Python

Learning Apache Spark with Python

- Docs »
- 6. Statistics and Linear Algebra Preliminaries
-

---

6. Statistics and Linear Algebra Preliminaries¶

Chinese proverb

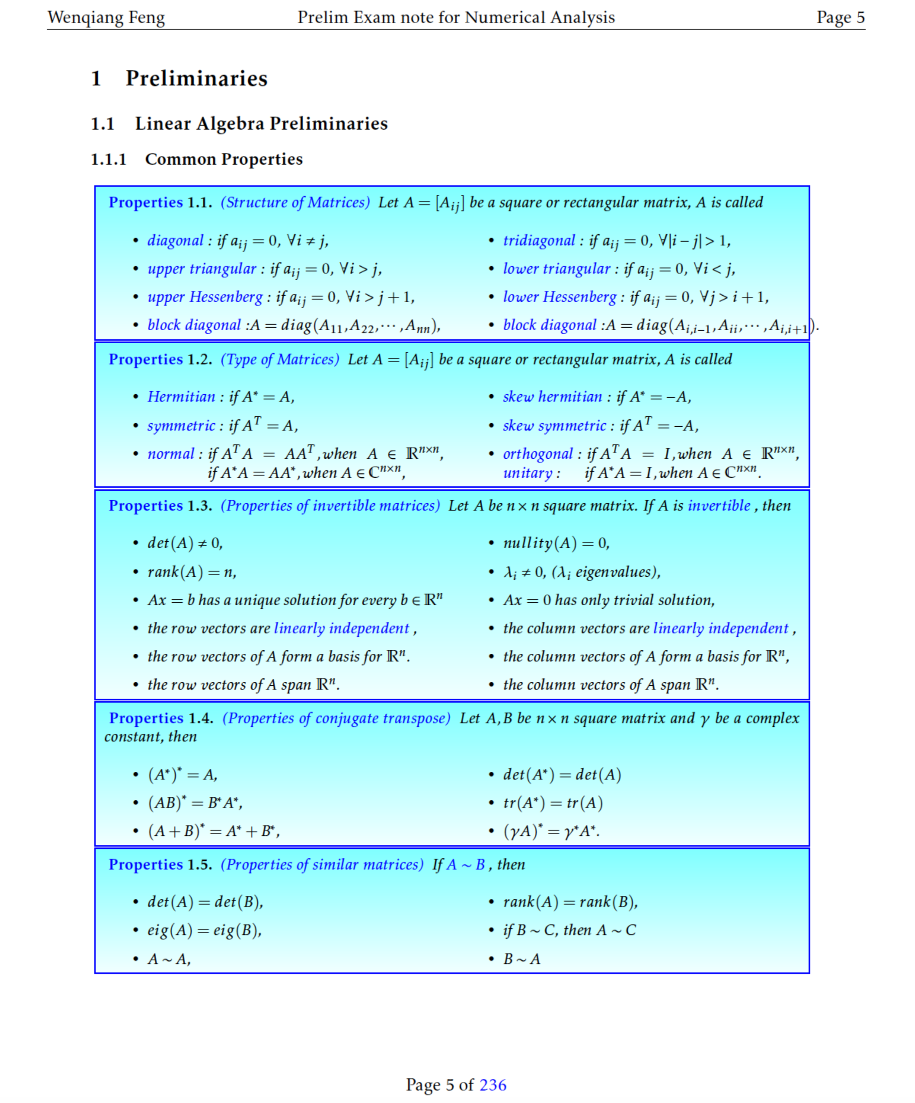
If you only know yourself, but not your opponent, you may win or may lose. If you know neither yourself nor your enemy, you will always endanger yourself – idiom, from Sunzi’s Art of War

6.1. Notations¶

- m : the number of the samples
- n : the number of the features
- y_i : i-th label
- $\hat{y}_i$: i-th predicted label
- $\bar{\mathbf{y}} = \frac{1}{m}\sum_{i=1}^{m} y_i$: the mean of $\mathbf{y}$.
- $\mathbf{y}$: the label vector.
- $\hat{\mathbf{y}}$: the predicted label vector.

6.2. Linear Algebra Preliminaries¶

Since I have documented the Linear Algebra Preliminaries in my Prelim Exam note for Numerical Analysis, the interested reader is referred to [Feng2014] for more details (Figure. Linear Algebra Preliminaries).



Wenqiang Feng Prelim Exam note for Numerical Analysis Page 5

1 Preliminaries

1.1 Linear Algebra Preliminaries

1.1.1 Common Properties

Properties 1.1. *(Structure of Matrices) Let $A = [A_{ij}]$ be a square or rectangular matrix, A is called*

- *diagonal : if $a_{ij} = 0, \forall i \neq j$,*
- *upper triangular : if $a_{ij} = 0, \forall i > j$,*
- *upper Hessenberg : if $a_{ij} = 0, \forall i > j+1$,*
- *block diagonal :$A = diag(A_{11}, A_{22}, \cdots, A_{nn})$,*
- *tridiagonal : if $a_{ij} = 0, \forall |i-j| > 1$,*
- *lower triangular : if $a_{ij} = 0, \forall i < j$,*
- *lower Hessenberg : if $a_{ij} = 0, \forall j > i+1$,*
- *block diagonal :$A = diag(A_{i,i-1}, A_{ii}, \cdots, A_{i,i+1})$.*

Properties 1.2. *(Type of Matrices) Let $A = [A_{ij}]$ be a square or rectangular matrix, A is called*

- *Hermitian : if $A^* = A$,*
- *symmetric : if $A^T = A$,*
- *normal : if $A^TA = AA^T$, when $A \in \mathbb{R}^{n\times n}$,*
  if $A^*A = AA^*$, when $A \in \mathbb{C}^{n\times n}$,
- *skew hermitian : if $A^* = -A$,*
- *skew symmetric : if $A^T = -A$,*
- *orthogonal : if $A^TA = I$, when $A \in \mathbb{R}^{n\times n}$,*
  unitary : if $A^*A = I$, when $A \in \mathbb{C}^{n\times n}$.

Properties 1.3. *(Properties of invertible matrices) Let A be $n \times n$ square matrix. If A is invertible , then*

- *$det(A) \neq 0$,*
- *$rank(A) = n$,*
- *$Ax = b$ has a unique solution for every $b \in \mathbb{R}^n$*
- *the row vectors are linearly independent ,*
- *the row vectors of A form a basis for $\mathbb{R}^n$.*
- *the row vectors of A span $\mathbb{R}^n$.*
- *$nullity(A) = 0$,*
- *$\lambda_i \neq 0$, (λ_i eigenvalues),*
- *$Ax = 0$ has only trivial solution,*
- *the column vectors are linearly independent ,*
- *the column vectors of A form a basis for $\mathbb{R}^n$,*
- *the column vectors of A span $\mathbb{R}^n$.*

Properties 1.4. *(Properties of conjugate transpose) Let A, B be $n \times n$ square matrix and γ be a complex constant, then*

- *$(A^*)^* = A$,*
- *$(AB)^* = B^*A^*$,*
- *$(A+B)^* = A^* + B^*$,*
- *$det(A^*) = det(A)$*
- *$tr(A^*) = tr(A)$*
- *$(\gamma A)^* = \gamma^* A^*$.*

Properties 1.5. *(Properties of similar matrices) If $A \sim B$, then*

- *$det(A) = det(B)$,*
- *$eig(A) = eig(B)$,*
- *$A \sim A$,*
- *$rank(A) = rank(B)$,*
- *if $B \sim C$, then $A \sim C$*
- *$B \sim A$*

Page 5 of 236

Linear Algebra Preliminaries¶

6.3. Measurement Formula¶

6.3.1. Mean absolute error¶

In statistics, **MAE** (Mean absolute error) is a measure of difference between two continuous variables. The Mean Absolute Error is given by:

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m} |\hat{y}_i - y_i|.$$

6.3.2. Mean squared error¶

In statistics, the **MSE** (Mean Squared Error) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors or deviations—that is, the difference between the estimator and what is estimated.

$$\text{MSE} = \frac{1}{m}\sum_{i=1}^{m} \left(\hat{y}_i - y_i\right)^2$$

6.3.3. Root Mean squared error¶

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{m}\sum_{i=1}^{m} \left(\hat{y}_i - y_i\right)^2}$$

6.3.4. Total sum of squares¶

In statistical data analysis the **TSS** (Total Sum of Squares) is a quantity that appears as part of a standard way of presenting results of such analyses. It is defined as being the sum, over all observations, of the squared differences of each observation from the overall mean.

$$\text{TSS} = \sum_{i=1}^{m} \left(y_i - \bar{\boldsymbol{y}}\right)^2$$

6.3.5. Explained Sum of Squares¶

In statistics, the **ESS** (Explained sum of squares), alternatively known as the model sum of squares or sum of squares due to regression.

The ESS is the sum of the squares of the differences of the predicted values and the mean value of the response variable which is given by:

$$\text{ESS} = \sum_{i=1}^{m} (\hat{y}_i - \bar{\boldsymbol{y}})^2$$

6.3.6. Residual Sum of Squares¶

In statistics, **RSS** (Residual sum of squares), also known as the sum of squared residuals (SSR) or the sum of squared errors of prediction (SSE), is the sum of the squares of residuals which is given by:

$$\text{RSS} = \sum_{i=1}^{m} (\hat{y}_i - y_i)^2$$

6.3.7. Coefficient of determination R^2¶

$$R^2 := \frac{ESS}{TSS} = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

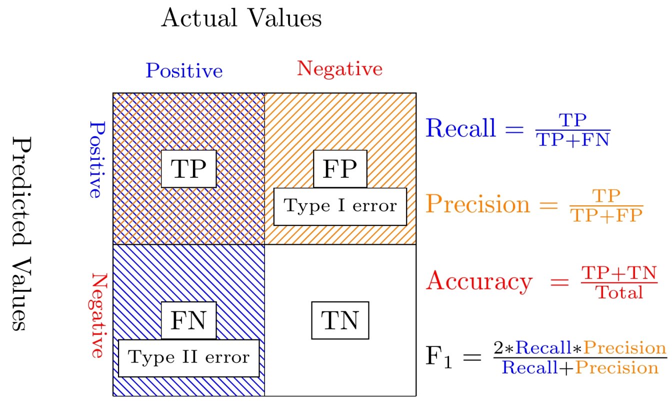
Note

In general ($\boldsymbol{y}^T\bar{\boldsymbol{y}} = \hat{\boldsymbol{y}}^T\bar{\boldsymbol{y}}$), total sum of squares = explained sum of squares + residual sum of squares, i.e.:

$$\text{TSS} = \text{ESS} + \text{RSS if and only if } \boldsymbol{y}^T\bar{\boldsymbol{y}} = \hat{\boldsymbol{y}}^T\bar{\boldsymbol{y}}.$$

More details can be found at Partitioning in the general ordinary least squares model.

6.4. Confusion Matrix¶


Actual Values
Positive
Negative
Predicted Values
Positive
Negative
TP
FP
Type I error
FN
Type II error
TN
Recall = $\frac{\text{TP}}{\text{TP+FN}}$
Precision = $\frac{\text{TP}}{\text{TP+FP}}$
Accuracy = $\frac{\text{TP+TN}}{\text{Total}}$
$\text{F}_1 = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall+Precision}}$

Confusion Matrix¶

6.4.1. Recall¶

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

6.4.2. Precision¶

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

6.4.3. Accuracy¶

$$\text{Accuracy} \ = \frac{\text{TP+TN}}{\text{Total}}$$

6.4.4. F_1-score¶

$$\text{F}_1 = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

6.5. Statistical Tests¶

6.5.1. Correlational Test¶

- Pearson correlation: Tests for the strength of the association between two continuous variables.
- Spearman correlation: Tests for the strength of the association between two ordinal variables (does not rely on the assumption of normal distributed data).
- Chi-square: Tests for the strength of the association between two categorical variables.

6.5.2. Comparison of Means test¶

- Paired T-test: Tests for difference between two related variables.
- Independent T-test: Tests for difference between two independent variables.
- ANOVA: Tests the difference between group means after any other variance in the outcome variable is accounted for.

6.5.3. Non-parametric Test¶

- Wilcoxon rank-sum test: Tests for difference between two independent variables - takes into account magnitude and direction of difference.

- Wilcoxon sign-rank test: Tests for difference between two related variables - takes into account magnitude and direction of difference.
- Sign test: Tests if two related variables are different – ignores magnitude of change, only takes into account direction.

Next Previous

© Copyright 2017, Wenqiang Feng Last updated on Dec 05, 2021.