A Statistics and Matrix Algebra Rosetta Stone

Overview

The knowledge that you already have in statistics can serve as an anchor for understanding concepts in matrix algebra and machine learning. Informative relationships can be found between covariance and inner products, beta and projection, and correlation and cosine distance. This is valuable, low-hanging fruit.

Below, we review basic statistics concepts from a more advanced perspective. Then we explore how those concepts can be represented with matrix algebra.

Statistics Review

Covariance

Covariance measures how much two variables co-vary. This is achieved by measuring the average of how much two variables deviate from their own means at the same time.

$$\mathrm{Cov}(X,\ Y)\ =\ \frac{1}{n}\sum_{i=1}^{n}(X_i\ -\ \bar{X})(Y_i\ -\ \bar{Y})$$

You will truly understand this when you realize that sorting two data columns, each from largest to smallest, maximizes their covariance. Furthermore, covariance, and therefore correlation, is a function of order; and we can shuffle data in order to impute a target level of correlation. That is the machinery of a copula.

R-Square and Correlation

The measure of covariance is in an awkward squared space. To make the measures of covariance interpretable, we can divide them by the total variance of X and Y.

$$R^2 = \frac{\mathrm{Cov}(X,Y)^2}{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}$$

The resulting proportion of the covariation of X and Y as a fraction of all the variation of X and Y is the R-square. We can square root this value to move back to our one-dimensional metrics, and the result is correlation.

$$\mathrm{Corr}(X,Y) = \sqrt{R^2}$$

We can summarize by saying that correlation is normalized covariance.

Variance and Standard Deviation

We can think of variance as the covariance of a variable with itself.

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X)$$

Variance measures the average squared deviation from the mean:

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

We can also think of variance as the original metric that measures dispersion as the average squared deviation. A consequence of that choice is that the arithmetic of risk will occur in squared space. For example, if you want to add measures of dispersion, you will add variances; and if you wanted to observe a fraction of explained dispersion, you will use ratios of variances. The act of square rooting these variances to calculate a standard deviation, moves us back to a one-dimensional metric and communicates the result.

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$$

Beta

Beta can be calculated by identifying the shared covariation between X and Y as a fraction of all the variation in X:

$$\beta = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}$$

Alternatively, Beta can be calculated starting with correlation, scaled by the ratio of standard deviations. Those standard deviations can be thought of as the relative step size for each of X and Y.

$$\beta = \mathrm{Corr}(X, Y) \cdot \frac{\mathrm{SD}(Y)}{\mathrm{SD}(X)}$$

Matrix Algebra

Centering Data

Before proceeding, the trick that unlocks the following dual interpretations is work withing centered data. That is, the means have been subtracted from the data columns. From here onward, we will refer to data columns as vectors and it is assumed that the vectors represent centered data columns.

$$\mathrm{X} = X^* - \bar{X}$$

Covariance, Correlation, and Beta

Covariance is proportional to the inner product of two vectors:

$$\mathrm{Cov}(\mathrm{X},\ \mathrm{Y}) = \frac{1}{\mathrm{n}}\ \langle X,\ \mathrm{Y}\ \rangle$$

The covariance matrix Σ for a multivariate dataset can be expressed as:

$$\Sigma = \frac{1}{n} X^T \mathrm{X}$$

Correlation and beta are derived as follows:

$$\mathrm{Corr}(X,Y) = \frac{\langle X,Y\rangle}{|X||Y|},\ \beta = \frac{\langle X,Y\rangle}{\langle X,X\rangle}$$

The norm of a vector corresponds to the standard deviation:

$$|\mathrm{X}| = \sqrt{\frac{1}{n}\langle X,X\rangle}$$

If the mean has been estimated from data, then a degree of freedom has been lost, and the sum of deviations should be unitized by dividing by n-1. See the appendix 'Where does that degree of freedom go?' for detail.

Projection

The projection of a vector v onto w is:

$$p = sw, \quad s = \frac{\langle v,w\rangle}{\langle w,w\rangle}$$

Cosine similarity between two vectors is:

$$\text{Cosine Similarity} = \frac{\langle X,Y\rangle}{|X||Y|}$$

This is equivalent to correlation for centered data vectors.

Projection onto a Line

Let's describe projection onto a line. We will do this twice, once in standard terminology and then with more applied terms.

Let's project vector $\boldsymbol{v}$ onto vector $\boldsymbol{w}$. We are looking for a projection vector \(\mathbf{p}\), as the location in $\boldsymbol{w}$ that is closest to $\boldsymbol{v}$. $\boldsymbol{p}$ is a scalar s multiple of $\boldsymbol{w}$.

$$\boldsymbol{p} = s\boldsymbol{w}$$

Let $\boldsymbol{r}$ represent the residual distance from $\boldsymbol{v}$ to $\boldsymbol{p}$. Visually, $\boldsymbol{r}$ drops down from $\boldsymbol{v}$ onto $\boldsymbol{p}$.

Mathematically, $\boldsymbol{v} + \boldsymbol{r} = \boldsymbol{p}$ and we choose $\boldsymbol{p}$ so that $\boldsymbol{r}$ falls onto it orthogonally, i.e., $\boldsymbol{r} \cdot \boldsymbol{p} = 0$.

Derivation

$$\boldsymbol{p} = s\boldsymbol{w}$$

$$\boldsymbol{v} + \boldsymbol{r} = \boldsymbol{p}$$

$$\boldsymbol{r} \cdot \boldsymbol{w} = 0$$

We solve for the scalar (s):

$$\boldsymbol{v} + \boldsymbol{r} = s\boldsymbol{w}$$

Multiply everything by $\boldsymbol{w}$:

$$\boldsymbol{v} \cdot \boldsymbol{w} + \boldsymbol{r} \cdot \boldsymbol{w} = s(\boldsymbol{w} \cdot \boldsymbol{w})$$

Since $\boldsymbol{r} \cdot \boldsymbol{w} = 0$, this simplifies to:

$$s = \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{\boldsymbol{w} \cdot \boldsymbol{w}}$$

Generalization to Subspaces

$\boldsymbol{w}$ does not need to be a vector; it could be a multi-dimensional subspace W.

Beta in Context

Starting with a data vector $\boldsymbol{y}$ and predictor vector $\boldsymbol{x}$, we want to project $\boldsymbol{y}$ onto $\beta\boldsymbol{x}$. Notice the difference; we are projecting $\boldsymbol{y}$ onto $\hat{\mathrm{y}}$ i.e. $\beta\boldsymbol{x}$. The math proceeds similarly after defining $\boldsymbol{r}$ as the residual vector:

$$\boldsymbol{y} + \boldsymbol{r} = \beta\boldsymbol{x}$$

$$\boldsymbol{x} \cdot \boldsymbol{y} + \boldsymbol{x} \cdot \boldsymbol{r} = \beta(\boldsymbol{x} \cdot \boldsymbol{x})$$

Since $\boldsymbol{x} \cdot \boldsymbol{r} = 0$:

$$\boldsymbol{x} \cdot \boldsymbol{y} = \beta(\boldsymbol{x} \cdot \boldsymbol{x})$$

$$\beta = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\boldsymbol{x} \cdot \boldsymbol{x}}$$

This schema generalizes to multiple linear regression:

$$\boldsymbol{y} + \boldsymbol{r} = \boldsymbol{X}\beta$$

$$\boldsymbol{X^T y} + \boldsymbol{X^T r} = \boldsymbol{X^T X \beta}$$

Since $\boldsymbol{X^T r} = 0$:

$$\boldsymbol{X^T y} = \boldsymbol{X^T X}\beta$$

$$\beta = (\boldsymbol{X^T X})^{-1}\boldsymbol{X^T y}$$

This time, we projected into the subspace of $\boldsymbol{X}$. We identified the subspace of $\boldsymbol{X}$ that is closest to $\boldsymbol{y}$. Following the principle that the shortest distance between two points is a straight line, we have identified the closest subspace of $\boldsymbol{X}$ by projecting onto it. This reasoning is what allows for this closed-form solution to Ordinary Least Squares.

Cosine Similarity and Correlation

Cosine similarity is used in natural language processing (NLP), recommendation systems, and data clustering. In NLP, it measures the similarity between texts, such as comparing word embeddings, sentences, or documents, for tasks like semantic search and text classification. In recommendation systems, cosine similarity helps identify users or items with similar preferences. It is also used in clustering algorithms, especially in high-dimensional and sparse datasets like social networks or document corpora.

Cosine is defined as:

$$\cos(\theta) = \frac{\text{adjacent}}{\text{hypotenuse}}$$

Replacing terminology with $\boldsymbol{x}\ and\ \boldsymbol{y}$:

$$\cos(\theta) = \frac{|s\boldsymbol{x}|}{|\boldsymbol{y}|}$$

Using the projection formula, $s = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\boldsymbol{x}\cdot\boldsymbol{x}}$, we derive:

$$\cos(\theta) = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|}$$

This shows how cosine similarity is equivalent to correlation for centered data.

Projection and Spatial Algorithms

Projection forms the basis for algorithms like Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM).

LDA Objective Function

$$\frac{(\text{mean projection for A} - \text{mean projection for B})^2}{\text{variance projection for A} + \text{variance projection for B}}$$

SVM

Identify $\boldsymbol{w}$ such that a hyperplane splits the categories.

Understanding these principles deepens comprehension of regression, LDA, and SVM.

Appendix: Where Does That Degree of Freedom Go?

If you know the population mean μ, then the formula for variance is:

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

This formula works because μ is a known constant. However, if $\bar{x}$ (the sample mean) is used instead of μ, there is a problem.

Specifically, $\bar{x}$ comes from the data and "uses up" some of the information. In fact, it uses $1/n$ of the data per data point. To demonstrate, consider this estimation of variance for three data points:

$$(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2$$

Do you see anything strange? Maybe not yet, but let's expand $\bar{x}$:

$$\bar{x} = \frac{x_1 + x_2 + x_3}{3}$$

Substituting this back, the expression becomes:

$$\left(x_1 - \frac{x_1 + x_2 + x_3}{3}\right)^2 + \left(x_2 - \frac{x_1 + x_2 + x_3}{3}\right)^2 + \left(x_3 - \frac{x_1 + x_2 + x_3}{3}\right)^2$$

Notice that for each of these measures of squared deviation, the data point x_i is going to subtract $1/3$ of itself and get a perfect 0 for part of the calculation. This 0 was supposed to be a measure of dispersion, and it would have been if we had been able to use μ.

The consequence of using $\bar{x}$ is that we lose $1/n$ of a measure of dispersion for each data point, and this happens n times. Ultimately, we lose one full unit of dispersion, also known as one degree of freedom.

This is why the formula for sample variance includes an adjustment:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

Here, $n - 1$ accounts for the lost degree of freedom due to the estimation of $\bar{x}$ from the data.