▶ Distancia de Mahalanobis.

- Es una medida de distancia. (Mahalanobis, 1936).

- Sirve para determinar la similitud entre 2 variables aleatorias multidimensionales.

- A diferencia de la distancia euclidea, esta tiene en cuenta la correlación entre las variables aleatorias.

$$\boxed{d_m(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T \, \Sigma^{-1} (\vec{x}-\vec{y})}}$$

▶ Distancia de Mahalanobis.
- $\vec{x}, \vec{y}$: variables aleatorias (con misma distribución de probabilidad.
- Σ^{-1} : matriz de covarianza.

Ejemplo: pescador mide la similitud entre 2 salmones, para venderlos + grandes, + caros.

Variables: anchura y longitud.

~~Vector~~ salmón i. construye un vector $\vec{x_i} = (\underset{\text{anchura}}{x_{1}}{}_{\underset{\text{salmón } i}{i}}, x_{2i})^T$

- longitud $\in [50, 200]$ cm
- anchura $\in [10, 20]$ cm

de probabilidad.

- Σ : matriz de covarianza.

Ejemplo: pescador mide la similitud entre 2 salmones, para venderlos + grandes, + caros.

Variables: anchura y longitud.

~~∀ vect.~~ salmón i construye un vector $\vec{x_i} = (x_{1i}, x_{2i})^T$

anchura

salmón i

- longitud $\in [50, 100]$ cm
- anchura $\in [10, 20]$ cm
- Con distancia euclídea $d_e(\vec{x_1}, \vec{x_2}) = \sqrt{(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2}$
  - La anchura varía menos que la longitud, entonces le da más importancia a la longitud. Entonces pondera según la varianza. Las variables con menos varianza tiene más importancia que las de mayor varianza. Entonces se iguala la importancia entre las variables.

$$d_2(\vec{x_1}, \vec{x_2}) = \left(\frac{(x_{11} - x_{12})^2}{\sigma_1} + \frac{(x_{21} - x_{22})^2}{\sigma_2} \right)^{1/2}$$

Desviación estándar para la longitud

σ_i: es la desviación estándar de la componente i de los vectores de medidas.

$$\boxed{d_e(\vec{x_1}, \vec{x_2}) = \sqrt{(\vec{x_1} - \vec{x_2})^T S^{-1} (\vec{x_1} - \vec{x_2})}}$$

▶ Anterior expresión en notación vectorial.

- S: matriz diagonal cuyos elementos en la diagonal son:

$$S_{ii} = \sigma_i^2$$

- La conchura depende de la longitud.

Entonces para tener en cuenta la dependencia de las 2 variables,

S se cambia por Σ: Matriz de covarianza.

$$\boxed{d_m(\vec{x_1}, \vec{x_2}) = \sqrt{(\vec{x_1} - \vec{x_2})^T \Sigma^{-1} (\vec{x_1} - \vec{x_2})}}$$

▶ Distancia de Mahalanobis

- Σ: Matriz de covarianza.

- Matriz de covarianza: Matriz □ que contiene la covarianza ~~datos~~ entre elementos de un vector.

$$\vec{X} = \begin{bmatrix} x_1 \\ \vdots \end{bmatrix}$$

▶ Vector aleatorio

La variable x_i es una aleatoria con varianza

S se cambia por Σ: Matriz de covarianza.

$$\boxed{d_m(\vec{x_1},\vec{x_2}) = \sqrt{(\vec{x_1}-\vec{x_2})^T \Sigma^{-1} (\vec{x_1}-\vec{x_2})}}$$;

▶ Distancia de Mahalanobis

- Σ: Matriz de covarianza.

- Matriz de covarianza: Matriz □ que contiene la covarianza ~~datos~~ entre elementos de un vector.

$$\boxed{\vec{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}}$$

▶ Vector aleatorio

La variable x_i es una aleatoria con varianza finita.

La matriz de covarianza será una $n \times n$ cuya entrada (i,j) es la covarianza entre la variable x_i y x_j:

$$\Sigma_{i,j} = \mathrm{Cov}(x_i, x_j)$$

- Covarianza: grado de variación conjunta de dos variables aleatorias respecto a sus medias.

~~Ayuda~~ Es lo que se necesita para determinar si dos variables son dependientes

▶ Criterio de Información Bayesiano (BIC).

Sirve para medir la calidad relativa entre varios modelos.
No dice que tan bueno es un modelo, sino, entre los que se tienen, cuál es el mejor.

$$\boxed{BIC = k\ln(n) - 2\ln(\hat{L})}$$

▶ Fórmula del criterio de información de Bayes.

- k: # de parámetros que tenga el modelo. (p para uno de regresión).
- $\hat{L}$: Función de Máxima Verosimilitud
- n: tamaño de la muestra, el # de datos.