Reducción de dimensionalidad y clasificación multiclase con Análisis Discriminante Lineal (LDA) robusto y no paramétrico

Abelino Sepúlveda, Henry Laniado
asepulvede@eafit.edu.co, hlaniado@eafit.edu.co

Escuela de ciencias aplicadas e ingeniería
Universidad EAFIT, Medellín, Colombia

22 de noviembre de 2022

Resumen. El Análisis Discriminante Lineal (LDA) es un método de reducción de dimensionalidad y clasificación supervisada, que muchos mucho de estos es sensible ante presencia de observaciones atípicas. Por lo que se propone e implementa una versión robusta de este método donde se interviene la matriz de covarianzas. Estas intervenciones se realizan por medio de los coeficientes de correlación de Kendall y Spearman y una implementación robusta del coeficiente de Pearson, adicional a esto, también se implementan los métodos basados en el encogimiento de matriz de covarianzas Shrinkages y el de determinante mínimo fastMCD. Estos métodos propuestos evidencian que tienen mejores desempaños que el de la librería de Python basado en valor singular de descomposición.

Keywords: Análisis Discriminante Lineal, Kendall, Pearson, Spearman, Shrinkages, fastMCD, robusto, no paramétrico.

1. Introducción

Hoy en día, la clasificación supervisada es una de las tareas que más se lleva a cabo en los modelos de aprendizaje automático (Singh *et al.*, 2016). Los cuales suelen usarse para problemas de clasificación y de regresión, que suelen distinguirse por el tipo de la variable

objetivo. Los principales usos de estos modelos son la identificación (Omar *et al.*, 2013), diagnóstico (Fatima *et al.*, 2017), detección y predicción (Sharma *et al.*, 2011) de variables del estudio de interés.

Entre los algoritmos más habituales de aprendizaje automático caracterizados por realizar tareas propias de clasificación se encuentran Árboles de Decisión, Clasificación de Naïve Bayes, Support Vector Machines (SVM), Métodos "Ensemble" y dentro de los paradigmas desarrollados por la estadística contamos con modelos de Regresión por mínimos cuadrados, Regresión Logística, Análisis Discriminante, entre otros. Estos algoritmos, son ampliamente utilizados gracias a su contribución y desempeño en distintas áreas de conocimiento. Dentro de los principales áreas y aplicaciones encontramos la quimioinformatica (Mitchell, 2014), economía (Mullainathan & Spiess, 2017), reconocimineto de caracteres (Bishop & Nasrabadi, 2006) y muchas otras más aplicaciones.

Este proyecto se centra en el desarrollo de Análisis Discriminante Lineal (LDA) usado para reducción de dimensionalidad y clasificación. La elección de este método es precisamente porque combina las dos técnicas anteriormente mencionadas las cuales brindan beneficios tales como reducción de tiempos de entrenamiento y aumento en general en el desempeño del algoritmo (Van Der Maaten *et al.*, 2009). Adicional a esto, como mencionan Pohar *et al.* (2004) LDA presenta ventajas sobre la regresión logística cuando contamos con escenarios con pocos datos o clases muy separadas.

Finalmente, uno de los principales problemas cuando trabajamos con algoritmos de clasificación es la presencia de datos atípicos también conocidos como outliers, ya que en muchas ocasiones estos tienden a sesgar nuestras estimaciones (Liano, 1996). Por esta razón, en la literatura se han propuesto muchos enfoques para combatir este problema, y en este paper se abarca algunos de estos enfoques para contrastar este problema, es decir, se va a trabajar una versión robusta de Análisis Discriminante Lineal.

2. Metodología

El Análisis Discriminante Lineal es un método de clasificación supervisado de variables cualitativas en el que dos o más grupos son conocidos a priori y nuevas observaciones se clasifican en uno de ellos en función de sus características (Izenman, 2008). Comúnmente conocido para la reducción de dimensiones antes de la clasificación.

Sea $\mathcal{P}$ una población particionada en k clases denotadas por $\Pi_1, \Pi_2, ..., \Pi_k$, donde cada ítem de la población está clasificada en una y solo una de las clases. Ahora, consideremos un r-vector $\mathbf{X}$ el cual representa el conjunto de observaciones a ser clasificados en las distintas clases, denotado por

$$\mathbf{X} = (X_1, X_2, ..., X_r)$$

2.1. LDA para clasificación binaria

Consideremos primero el problema de clasificación binaria ($k = 2$). Queremos discriminar entre las clases Π_1, Π_2.

2.1.1. Clasificador de la regla de Bayes

Sea

$$P(\mathbf{X} \in \Pi_i) = \pi_i \quad i = 1, 2 \tag{1}$$

Las probabilidades a priori al ser seleccionada aleatoriamente una observación **X=x** pertenezca a Π_1 o Π_2. Supongamos también que la densidad de probabilidad multivariada condicional de $\mathbf{X}$ para la i-ésima clase es

$$\mathrm{P}\left(\mathbf{X} = \mathbf{x} \mid \mathbf{X} \in \Pi_i\right) = f_i(\mathbf{x}), \quad i = 1, 2 \tag{2}$$

Notemos que no hay requisitos de que f_i tiene que ser continuo; puede ser discreto o una mixtura finita de distribuciones. Ahora, de las Ecuaciones 1 y 2 teorema de Bayes produce la probabilidad posterior

$$p\left(\Pi_i \mid \mathbf{x}\right) = \mathrm{P}\left(\mathbf{X} \in \Pi_i \mid \mathbf{X} = \mathbf{x}\right) = \frac{f_i(\mathbf{x})\pi_i}{f_1(\mathbf{x})\pi_1 + f_2(\mathbf{x})\pi_2} \tag{3}$$

de que la observación $\mathbf{x}$ pertenezca a Π_i, $i = 1, 2$.

Para una $\mathbf{x}$ dada, una estrategia de clasificación razonable es asignar $\mathbf{x}$ a esa clase con la probabilidad posterior más alta. Esta estrategia es llamada *Clasificador de la regla de Bayes*. El cual indica que la observación $\mathbf{x}$ es asignada a Π_1 si

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{\pi_2}{\pi_1} \tag{4}$$

y es asignado a Π_2 en caso contrario.

2.1.2. Análisis Discriminante Lineal

Ahora hacemos que el clasificador de la regla de Bayes sea más específico sin seguir suposiciones como distribución normal de clases o covarianzas iguales entre las clases (Ye, 2007). Consideremos dos clases de observaciones con medias $\vec{\mu}_1, \vec{\mu}_2$ y covarianzas $\mathbf{\Sigma_1}, \mathbf{\Sigma_2}$. Fisher definió la separación entre estas dos distribuciones por la proporción de la varianza entre las clases, entre la varianza dentro de las clases

$$S = \frac{(\vec{w} \cdot \vec{\mu}_2 - \vec{w} \cdot \vec{\mu}_1)^2}{\vec{w}^T \Sigma_1 \vec{w} + \vec{w}^T \Sigma_0 \vec{w}} = \frac{(\vec{w} \cdot (\vec{\mu}_2 - \vec{\mu}_1))^2}{\vec{w}^T (\Sigma_0 + \Sigma_1) \vec{w}} \tag{5}$$

y la separación máxima de las clases ocurre cuando

$$\vec{w} = (\Sigma_1 + \Sigma_2)^{-1} (\vec{\mu}_2 - \vec{\mu}_1) \tag{6}$$

y definimos una función lineal

$$L(\mathbf{x}) = w_0 + \vec{w}^T \mathbf{x} \tag{7}$$

donde

$$w_0 = -\frac{1}{2} \left\{ \vec{\mu}_1^{\tau} (\Sigma_1 + \Sigma_2)^{-1} \vec{\mu}_1 - \vec{\mu}_2^{\tau} (\Sigma_1 + \Sigma_2)^{-1} \vec{\mu}_2 \right\} + \log_e (\pi_2 / \pi_1) \tag{8}$$

Asignamos a $\mathbf{x}$ en Π_1 si $L(\mathbf{x}) > 0$ y en Π_2 en el caso contrario.

2.2. LDA para clasificación multiclase

Considerando un problema de clasificación multiclase ($k = c$). Queremos determinar entre las clases $\Pi_1, \Pi_2, ..., \Pi_c$. Sean $\vec{\mu}_i$ y Σ_i, $i = 1, 2, ..., c$ el vector de medias y la matriz de covarianzas de cada una de las clases. Definimos las matrices de covarianzas *within* y *between* de la siguiente forma

$$\begin{aligned} \Sigma_{within} &= \sum_{i=1}^{c} \Sigma_i \\ \Sigma_{between} &= \frac{1}{c} \sum_{i=1}^{c} (\vec{\mu}_i - \vec{\mu}) (\vec{\mu}_i - \vec{\mu})^T \end{aligned} \tag{9}$$

donde $\vec{\mu}$ es la media de las medias de las clases. Ahora, definimos

$$\vec{L} = \log (p_i) - \frac{1}{2} \log (\det (\Sigma_i)) \frac{1}{2} \vec{\mu}_i^{\tau} \Sigma_i^{-1} \vec{\mu}_i + \vec{\mu}_i^{\tau} \Sigma_i^{-1} \mathbf{x} - \frac{1}{2} \mathbf{x}^{\tau} \Sigma_i^{-1} \mathbf{x} \tag{10}$$

el c-vector que define a qué clase se asigna $\mathbf{x}$. Este criterio de asignación consiste de la siguiente forma: se asigna $\mathbf{x}$ a la clase para el cual L_i, $i = 1, 2, ..., c$ es mayor.

2.3. LDA para reducción de dimensionalidad

Como se mencionó anteriormente, una de las principales características de LDA es que además de clasificar también reduce dimensionalidad. Definimos una matriz W de la siguiente forma

$$W = \Sigma_{within}^{-1}\Sigma_{between} \tag{11}$$

Y calculamos los vectores y valores propios resolviendo el siguiente esquema

$$(W - \lambda I) = 0$$

2.3.1. Varianza explicada

Una vez calculados los vectores y valores propios con la Ecuación 11 se puede calcular el porcentaje de varianza explicada por cada unos de los features.

$$\text{Varianza Explicada} = \frac{\lambda_i}{\sum_{i \in d} \lambda_i} \quad i = 1, 2, ..., d$$

donde d es el numero de columnas o features. Con este porcentaje podemos elegir las variables que mejor discriminen los datos para obtener mejor predicciones.

2.4. LDA Robusto

La presencia de valores atípicos u outliers en los conjunto de datos son cada vez más comunes, pudiendo distorsionar seriamente el comportamiento de los contrastes estadısticos en las muestras o poblaciones (Ghosh & Vogt, 2012). Por este motivo, la propuesta a metodologías para estimaciones robustas son cada vez más usadas gracias a las ventajas que estas generan. Una de estas metodologías para contrarrestar los outliers es por medio de las estimaciones robustas de la matriz de covarianzas (Little, 1988). Esta estimación se ha convertido en un amplio objeto de estudio. Del cual, este proyecto se enfocará en cuatro de estas estimaciones:

2.4.1. Coeficientes de correlación

coeficiente	Kendall	Spearman	Pearson robusto
$\hat{cov}(x, y)$	$\rho_k * \hat{\sigma}(x) * \hat{\sigma}(y)$	$\rho_s * \hat{\sigma}(x) * \hat{\sigma}(y)$	$\rho_r * \hat{\sigma}(x) * \hat{\sigma}(y)$

Cuadro 1: Estimación robusta de la matriz de covarianzas por coeficientes de correlación

donde

$$\rho_k(x, y) = \frac{2(N_c - N_d)}{n(n-1)}; \;\; \rho_s(x, y) = 1 - \frac{6\sum d_i^2}{n(n^2-1)}; \;\; \rho_r(x, y) = \frac{C_r(x, y)}{\sqrt{C_r(x, x)C_r(y, y)}} \tag{12}$$

con N_c como el número de pare concordantes, N_d pares discordantes, n el número de observaciones, d_i la diferencia entre los dos rangos de cada observación y finalmente,

$$\begin{aligned} \hat{\sigma}(x) &= median\left\{|x_i - median(x)|\right\} \\ C_r(x,y) &= median\left\{(x_i - median\left(x_i\right))\left(x_i - median\left(x_i\right)\right)\right\} \end{aligned} \tag{13}$$

2.4.2. Shrinkages

Técnica basada en el encogimiento de la matriz de covarianzas el cual tiene como objetivo la compensasión entre el sesgo y la varianza (Gutiérrez-Sepúlveda *et al.*, 2018).

2.4.3. fastMCD

El método de determinante mínimo de la matriz de covarianza, fue propuesto por Rousseauw en 1984 cuyo objetivo es encontrar h observaciones (de N) cuya matriz de covarianza tenga el determinante más bajo (Rousseeuw & Driessen, 1999).

3. Resultados

Para el proceso de validación y testeo de la metodología, se realizaron pruebas de reducción de dimensionalidad y de clasificación con 4 dataset diferentes

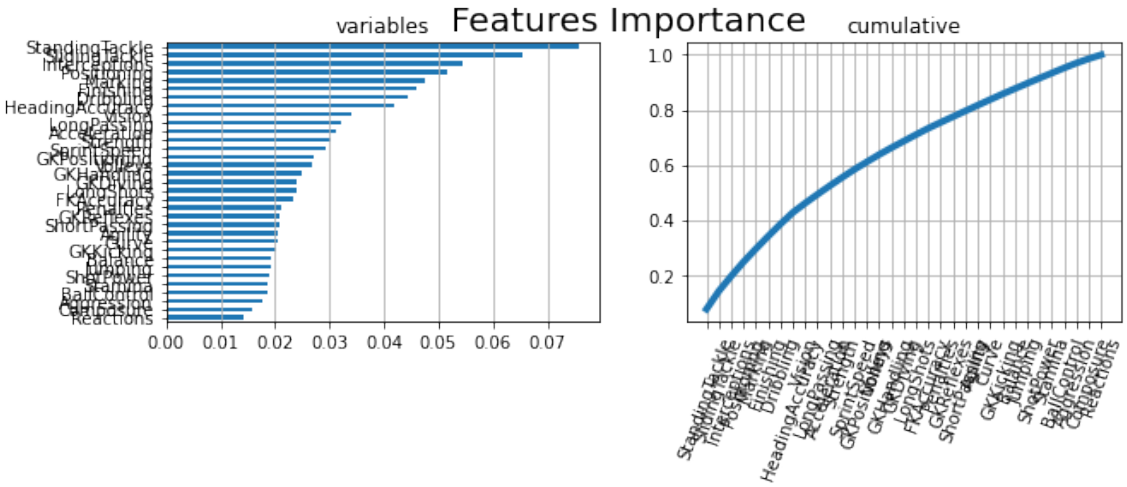
dataset	numero de filas	numero de columnas	numero de clases
FIFA	18147	33	4
iris setosa	150	4	3
wine	178	13	3

Cuadro 2: Datasets usados para testeo

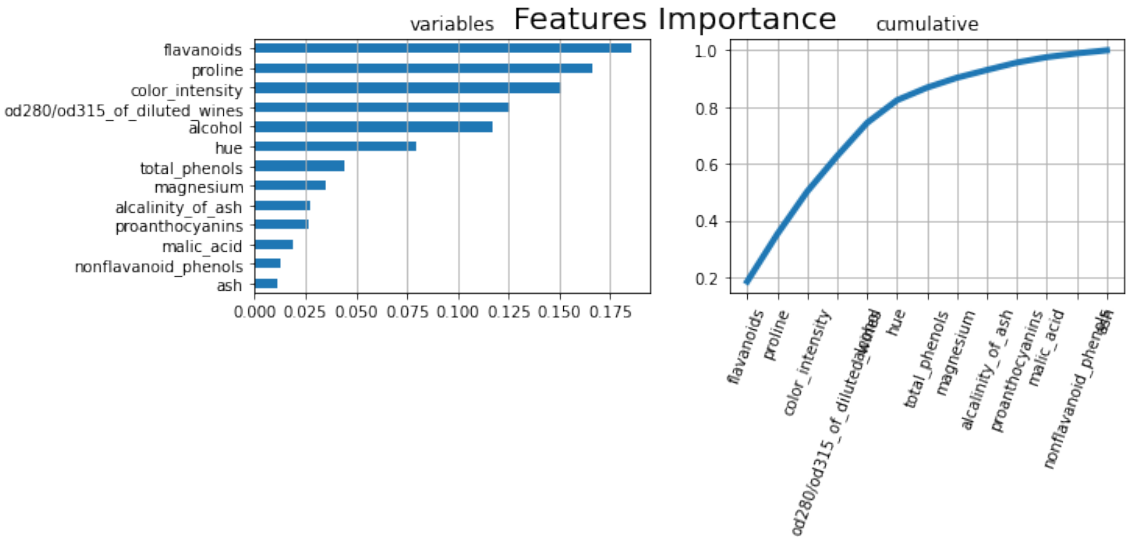
La Tabla 2 muestra los datasets utilizados, donde cada columna hace referencia al número de filas, número de columnas y el número de clases, donde podemos ver que el dataset de FIFA es que el mayor número de observaciones tiene por lo que sería el que resultados más interesantes podría generar.

3.0.1. Reducción de dimensionalidad

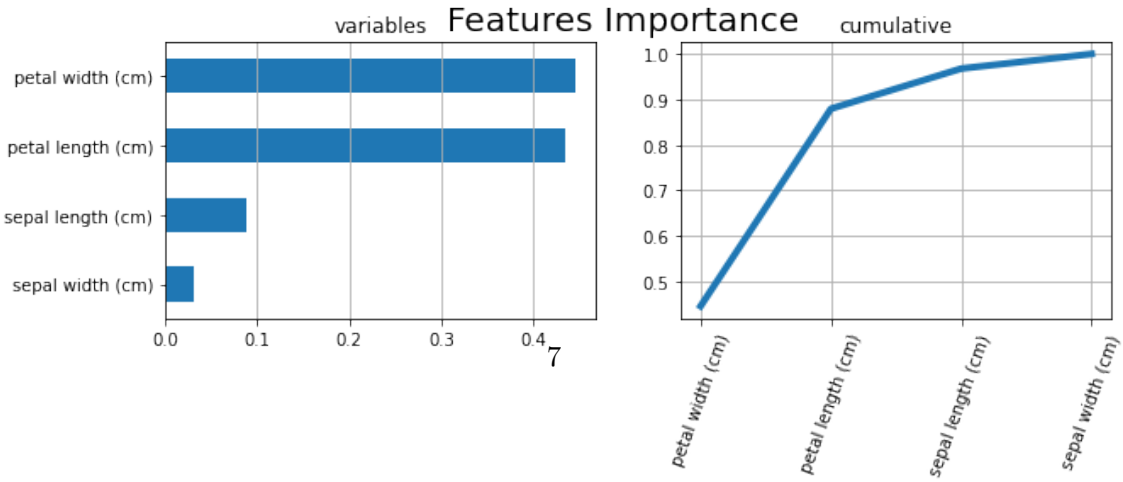
Inicialmente, se realizó el procedimiento de reducción de dimensionalidad. Donde se encuentra el porcentaje de varianza explicada de cada una de las variables para escoger las más explicativas para el proceso de clasificación.



((a)) Varianza explicada FIFA



((b)) Varianza explicada wine



((c)) Varianza explicada iris setosa

Figura 1: Varianza Explicada datasets

En la Figura 1 vemos el porcentaje de varianza explicado para cada una de las variables de cada dataset. Una observación particular es que en la Figura 1(a) el porcentaje de varianza explicado para cada variable no supera el 0.1, es decir, ninguna variable explica lo suficiente, por lo que tocaría considerar muchas variable para poder explicar por lo menos el 0.8 de los datos. Algo muy parecido ocurre con la Figura 2(b). Sin embargo, ocurre lo contrario en la Figura 1(c) donde con sólo las dos primeras variables ya tenemos un 0.8 de la varianza explicada de los datos.

Una vez elegidos las variables en base a la figura anterior, procedemos a realizar el procedimiento de clasificación.

3.0.2. Clasificación

Para este proceso, no se realizó ningún tipo de procesamiento de los datos, es decir, no se hizo ningún tipo de limpieza ya que el principal objetivo es evaluar el desempeño de estos algoritmos con presencia de outliers.

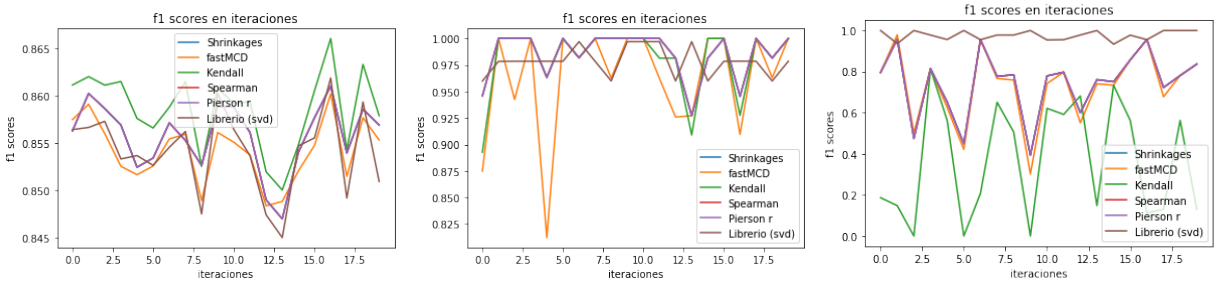
A continuación se muestra la tabla con los resultados del desempeño de cada uno de los métodos. Se realizaron 20 iteraciones, donde los datos eran separados aleatoriamente en cada iteración en muestras de entrenamiento y testeo. En cada iteración se calculaba el F1 score de cada uno de los métodos y finalmente se calculó un promedio de estos. Esto fue realizado de esta manera para mostrar si había consistencia en los resultados sin importar cómo estaba distribuidos los datos de testeo y entrenamiento.

metodo	Libreria (svd)	Pearson r	Kendall	Spearman	Shrinkages	fastMCD
FIFA	0.852	0.855	0.858	0.855	0.855	0.854
iris setosa	0.975	0.744	0.366	0.744	0.744	0.728
wine	0.982	0.985	0.980	0.985	0.985	0.963

Cuadro 3: Comparación F1 scores de los métodos

En la Tabla 3 observamos el F1 score de cada método para cada dataset. Con los datasets de FIFA y wine podemos ver cómo casi todo tienen el mismo desempeño, el cual es muy bueno para todos. Para el conjunto de datos de FIFA vemos cómo Kendall es quien tiene mejor desempeño, mientras que en wine Pearson robustto, Spearman y Shrinkages son quienes mejor desempeño llevan. Sin embargo, observamos que en Iris setosa, vemos que los métodos robustos tienen un mal desempeño en comparación al de la librería (*Singular value decomposition*). Esto es devido a que la matriz de covarianzas no era de rango completo, lo cual entorpece el proceso, ya que se necesita calcular la inversa de esta matriz.

Finalmente, se quiso realizar una visualización de la precisión en la clasificación en cada una de las iteraciones que se realizaron.



((a)) Scores FIFA ((b)) Scores wine ((c)) Scores iris setosa

Figura 2: F1 scores datasets en iteraciones

En la Figura 2 vemos la variación del F1 score en cada una de las iteraciones, en las Figuras 2(a) y 2(b) vemos que los métodos robustos tienen mejores desempeño que el de la librería. Mientras que en la Figura 2(c) vemos cómo siempre la librería es quien tiene mejor desempeño, esto posiblemente por las razones anteriormente mencionadas.

4. Conclusiones y futuro trabajo

En este artículo se estudia el Análisis Discriminante Lineal, el cual es un método de reducción y clasificación. Se proponen implementaciones robustas para intervenir la matriz de covarianzas y hacer que al método no tan sensible ante registros atípicos. Estas propuestas e implementaciones se evidencian que tienen buen desempeño.

Como direcciones de investigación futura se propone implementar una aproximación de la matriz inversa ya sea por Moore-Penrose para solucionar problemas de rangos no completos y finalmente, también se propone usar la propuesta original de reducción de dimensionalidad, es decir, no escoger las variables que mayormente expliquen la variabilidad si no, realizar una transformación en la matriz de datos para analizar si así se obtienen mejores resultados.

Referencias

Bishop, Christopher M, & Nasrabadi, Nasser M. 2006. *Pattern recognition and machine learning*. Vol. 4. Springer.

Fatima, Meherwar, Pasha, Maruf, *et al.* 2017. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, **9**(01), 1.

Ghosh, Dhiren, & Vogt, Andrew. 2012. Outliers: An evaluation of methodologies. *In: Joint statistical meetings*, vol. 2012.

Gutiérrez-Sepúlveda, Daniela, Laniado, Henry, & Medina-Hurtado, Santiago. 2018. Estimación robusta de la matriz de covarianza para la selección óptima de portafolios de inversión. *DYNA*, **85**(207), 328–336.

Izenman, Alan Julian. 2008. Modern multivariate statistical techniques. *Regression, classification and manifold learning*, **10**, 978–0.

Liano, Kadir. 1996. Robust error measure for supervised neural network learning with outliers. *IEEE Transactions on Neural Networks*, **7**(1), 246–250.

Little, Roderick JA. 1988. Robust estimation of the mean and covariance matrix from data with missing values. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **37**(1), 23–38.

Mitchell, John BO. 2014. Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, **4**(5), 468–481.

Mullainathan, Sendhil, & Spiess, Jann. 2017. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, **31**(2), 87–106.

Omar, Salima, Ngadi, Asri, & Jebur, Hamid H. 2013. Machine learning techniques for anomaly detection: an overview. *International Journal of Computer Applications*, **79**(2).

Pohar, Maja, Blas, Mateja, & Turk, Sandra. 2004. Comparison of logistic regression and linear discriminant analysis: a simulation study. *Metodoloski zvezki*, **1**(1), 143.

Rousseeuw, Peter J, & Driessen, Katrien Van. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**(3), 212–223.

Sharma, Navin, Sharma, Pranshu, Irwin, David, & Shenoy, Prashant. 2011. Predicting solar generation from weather forecasts using machine learning. *Pages 528–533 of: 2011 IEEE international conference on smart grid communications (SmartGridComm)*. IEEE.

Singh, Amanpreet, Thakur, Narina, & Sharma, Aakanksha. 2016. A review of supervised machine learning algorithms. *Pages 1310–1315 of: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. Ieee.

Van Der Maaten, Laurens, Postma, Eric, Van den Herik, Jaap, *et al.* 2009. Dimensionality reduction: a comparative. *J Mach Learn Res*, **10**(66-71), 13.

Ye, Jieping. 2007. Least squares linear discriminant analysis. *Pages 1087–1093 of: Proceedings of the 24th international conference on Machine learning.*