

TP2: Temas Avanzados

Entrega: Miércoles 21/11 (23:59hs)

Estar atentos a la sección “nuevos” al final del documento.

Para este trabajo, cada grupo deberá escribir un informe que explique un artículo del listado de abajo. Se espera que el informe contenga una explicación del tema con el nivel necesario como para introducir el tema a un estudiante de esta materia que no haya leído el artículo. No se espera que el informe sea autocontenido: puede dar por sabidos los temas vistos en clase, sin necesidad de explicarlos.

El informe deberá consistir en un archivo en formato PDF de tres páginas como máximo, con tamaño de fuente 10pt o mayor. La longitud del informe es parte de la evaluación: se penalizarán a los informes que excedan la longitud máxima.

En general, el contenido del informe (textos, figuras y tablas) debe ser original de los integrantes del grupo del TP. Si consideran necesario copiar algo textualmente, deben aclarar con claridad qué se copia y cuál es la fuente (paper/libro y número de página).

Elección del tema

El miércoles 07/11 solicitaremos la lista de exactamente tres (3) temas elegidos, indicando el orden de preferencia. También pueden sugerirse otros temas de interés y otros papers. En base a estas elecciones, se asignarán los temas tratando de respetar al máximo las preferencias.

Temas

Generales

- 1) Breiman, L. (2001). **Statistical Modeling: The Two Cultures**. Statistical Science (Vol. 16).
- 2) Zhang, M. L., & Zhou, Z. H. (2014). **A review on multi-label learning algorithms**. *IEEE transactions on knowledge and data engineering*, 26(8), 1819-1837.

Time series

- 3) Schäfer, P. (2014). **The BOSS is concerned with time series classification**.
- 4) Esling, P., & Agon, C. (2012). **Time-series data mining**. *ACM Computing Surveys*, 45(1), 1–34.

Detección de anomalías

- 5) Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). **Isolation forest**. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 413-422). IEEE.

Métodos de imputación

- 6) Stekhoven, D. J., & Bühlmann, P. (2011). **MissForest—non-parametric missing value imputation for mixed-type data**. *Bioinformatics*, 28(1), 112-118.

Feature selection

- 7) Perthame, É., Friguet, C., & Causeur, D. (2016). **Stability of feature selection in classification issues for high-dimensional correlated data.** *Statistics and Computing*, 26(4), 783–796.

Reducción Dimensional

- 8) Maaten, L. V. D., & Hinton, G. (2008). **Visualizing data using t-SNE.** *Journal of machine learning research*, 9(Nov), 2579-2605.

Clustering

- 9) Cheng, Y. (1995). **Mean shift, mode seeking, and clustering.** *IEEE transactions on pattern analysis and machine intelligence*, 17(8), 790-799.
- 10) Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). **On spectral clustering: Analysis and an algorithm.** In *Advances in neural information processing systems* (pp. 849-856).

Modelos Gráficos

- 11) Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). **Latent dirichlet allocation.** *Journal of machine Learning research*, 3(Jan), 993-1022.
- 12) Rabiner, L. R. (1989). **A tutorial on hidden Markov models and selected applications in speech recognition.** *Proceedings of the IEEE*, 77(2), 257-286.
- 13) Lafferty, J., McCallum, A., & Pereira, F. C. (2001). **Conditional random fields: Probabilistic models for segmenting and labeling sequence data.**

Interpretability

- 14) Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). **Why should i trust you?: Explaining the predictions of any classifier.** In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144). ACM.
- 15) Lundberg, S. M., & Lee, S. I. (2017). **A unified approach to interpreting model predictions.** In *Advances in Neural Information Processing Systems* (pp. 4765-4774).
- 16) Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018). **Manipulating and measuring model interpretability.** arXiv preprint arXiv:1802.07810.

KNN

- 17) Muja, M., & Lowe, D. G. (2014). **Scalable nearest neighbor algorithms for high dimensional data.** *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11), 2227-2240.

LDA

- 18) Prince, S. J., & Elder, J. H. (2007, October). **Probabilistic linear discriminant analysis for inferences about identity.** In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on* (pp. 1-8). IEEE.

GMMs

- 19) Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). **Speaker verification using adapted Gaussian mixture models.** *Digital signal processing*, 10(1-3), 19-41.

- 20) Leggetter, C. J., & Woodland, P. C. (1995). **Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models.** *Computer speech & language*, 9(2), 171-185.
- 21) Gauvain, J. L., & Lee, C. H. (1994). **Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains.** *IEEE transactions on speech and audio processing*, 2(2), 291-298.

NB

- 22) Zhang, H. (2004). **The optimality of naive Bayes.** *AA*, 1(2), 3.
- 23) Frank, E., Trigg, L., Holmes, G., & Witten, I. H. (2000). **Naive Bayes for regression.** *Machine Learning*, 41(1), 5-25.

Ensamblables

- 24) Kazemi, V., & Sullivan, J. (2014). **One millisecond face alignment with an ensemble of regression trees.** In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*(pp. 1867-1874).
- 25) Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). **Do we need hundreds of classifiers to solve real world classification problems?.** *The Journal of Machine Learning Research*, 15(1), 3133-3181.

Modelos Generativos vs. Discriminativos

- 26) Ng, A. Y., & Jordan, M. I. (2002). **On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes.** In *Advances in neural information processing systems* (pp. 841-848). + Xue, J. H., & Titterton, D. M. (2008). **Comment on “On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes”.** *Neural processing letters*, 28(3), 169.

Regularización

- 27) Hoerl, A. E., & Kennard, R. W. (1970). **Ridge regression: Biased estimation for nonorthogonal problems.** *Technometrics*, 12(1), 55-67.
- 28) Tibshirani, R. (1996). **Regression shrinkage and selection via the lasso.** *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.
- 29) Zou, H., & Hastie, T. (2005). **Regularization and variable selection via the elastic net.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320.

Nuevos

- 30) Ledoit, O., & Wolf, M. (2003). **Honey, I shrunk the sample covariance matrix.** (y comentar además cómo afecta a LDA)
- 31) Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). **Leakage in data mining: Formulation, detection, and avoidance.** *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4), 15.