# DUDAS Y ANOTACIONES

## DUDAS

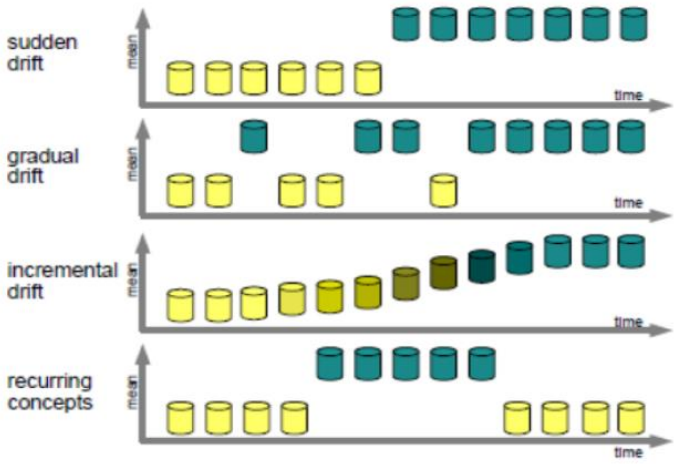
- Preguntar si es necesario meter los tipos de clustering *density-based*, *grid-based* y *model-based* o con *partitioning* y *micro-clusters*. -> Centrarnos en partitioning, hierarchical, model-based (probablistic) y micro-clusters.
    - **Partitioning algorithms:** construct a partition of a set of objects into k clusters, that minimize an objective function (e.g. the sum of squares distances to the centroid representative). Examples include k-means (Farnstrom et al., 2000), and k-medoids;
    - **Hierarchical algorithms.** (En otro sitio sale este tipo en lugar de micro-clustering)
    - **Micro-clustering algorithms:** divide the clustering process into two phases, where the first phase is online and summarizes the data stream in local models (micro-clusters) and the second phase generates a global cluster model from the micro-clusters. Examples of these algorithms include BIRCH (Zhang et al., 1996) and CluStream (Aggarwal et al., 2003);
    - **Density-based algorithms** are based on connectivity between regions and density functions. This type of algorithms find clusters of arbitrary shapes, e.g., DBSCAN (Birant and Kut, 2007), and OPTICS (Peter Kriegel et al., 2003). Se basa en la detección de en qué áreas existen concentraciones de puntos y dónde están separados por áreas vacías o con escasos puntos. Los puntos que no forman parte de un clúster se etiquetan como ruido.
    - **Grid-based algorithms:** based on a multiple-level granularity structure. View instance space as grid structures, e.g., Fractal Clustering (Barbará and Chen, 2000), and STING (Hinneburg and Keim, 1999). Grid based methods quantize the object space into a finite number of cells (hyper-rectangles) and then perform the required operations on the quantized space.
    - **Model-based algorithms:** find the best fit of the model to all the clusters. Good for conceptual clustering, e.g., COBWEB (Fisher, 1987), and SOM (Kaski and Kohonen, 1994). Model based clustering operates on the assumption that gene expression data originates from a finite mixture of underlying probability distributions (Ramoni et al. 2001). Each cluster corresponds to a different distribution, and generally, the distributions are assumed to be Gaussians.

- STREAM -> No sé que relación hay entre SSQ minimization y facility location -> ¿… and instead evaluates an algorithm's performance by a combination of SSQ and the numbers of centers used?
- STREAM -> k-Median : Nº of medians to be at most k; Facility location -> Range of number of centers. ¿Why facility location is more convinient if they

look for a z that gives a certain k? -> ¿K-Median es NP-Hard pero utilizando Facility Location en el Local Search es factible?

- ¿El recurring concept drift de la propuesta RGNBC (Redes Bayesianas) tiene que ver con la definición de este concepto? -> En RGNBC puede que no tenga que ver con un concept drift recurrente puesto que habla sobre la llegada de nuevos atributos.
- ¿Por qué en la ecuación 9 se tienen en cuenta aquellos valores de instancias que no pertenecen a la clase a? (Propuesta RGBNC)
- En la propuesta RGNBC, en la ecuación 12, ¿la evidencia no debería ser el sumatorio del numerador de posterior(…) en lugar del sumatorio de posterior(…)? -> Para calcular cada uno de los posterior necesitas la evidencia que se calcula con la suma de todos los posterior.
- Paper "A survey on rough set theory and its applications" -> ¿BND(X) and NEG(X) definitions interchanged? Además, duda con respecto a la utilidad que tiene el paper RGNBC
- Paper "RGNBC" -> Página 5 último párrafo -> Debería de hablar sobre la varianza en lugar de sobre la media.
- ¿Error en la gráfica de la página 5 del paper "Classification of Massive …"? ¿En la segunda también?
- Comentarle a los profesores lo de los diferentes surveys encontrados. Ejemplo: Ensemble (2017).
- No se entiende la función de CFIT en el paper "An effective pattern-based Bayesian classifier for evolving data stream". Utilización: último párrafo de la página 5.
- Paper "Mining Complex Models from Arbitrarly Large Databases in Constant Time" -> Difficult to understand.
- Preguntar si es necesario que los métodos de clasificación sean también multietiqueta. -> Si
- Preguntar si también interesan papers que traten las redes bayesianas en general para flujos de datos, sin ser clasificadores bayesianos explícitamente.
- Preguntar por el paper que contiene el método denominado Globally Adaptive-MB-MBC (supervisado multidimensional) -> No lo podemos encontrar -> ¿Tiene un paper publicado?
- Método CPL-DS (semi-supervisado unidimensional) -> ¿tiene un paper publicado? -> ¿Classifying evolving data streams with partially labeled data?
- Diferencia entre los artículos de la sección Refereed journals y los de la sección Conference and workshop
- Types of concept drift:
    - Real concept drift (change of the target concept that the classifier is trying to predict) and virtual concept drift (change of the underlying data distribution).
    - According to [14], there are three possible sources of concept drifts: Conditional change (real concept drift), feature change (virtual concept drift) and dual change.

- Sudden drift, gradual drift, incremental drift, recurring concepts -> ¿Rate of change?




Tenerlo en cuenta a la hora de hablar de concept drift

- En la página 4 del paper "Classifying evolving data streams with partially labeled data (2011)" -> otherwise, the window size increases to include the more recent instances. -> ¿No debería ser to include the more out-of-date instances?
- Paper "Classifying evolving data streams with partially labeled data (2011)" -> Página 7 -> ¿Por qué se muestrea solo de la distribución empírica del instante de tiempo s y no también del instante de tiempo s+1?
- Preguntar por las propuestas multi-etiqueta y multi-dimensional. Si hay que nombrarlas. -> Si
- Preguntar si me centro en los papers más recientes.
- Paper "Learning Decision Trees from Data Streams with Concept Drift (2016)" -> CEVOT extends EVO-Tree. This EVO-Tree has one split value in each node. ¿Why not more?

## ANOTACIONES

- Mirar libro para ver las relaciones entre métodos de clustering.
- Ver si el Sampling to Obtain Feasible Centers (apartado) de STREAM se puede comparar con el coreset nombrado en STREAMKM++.
- Comprobar si la carpeta Dynamic, Temporal and Continuous Time Bayesian Networks se puede incluir dentro de la carpeta de Bayesian networks que está dentro de la carpeta Classification for data streams.
- Mirar de las propuestas de clasificadores bayesianos a la hora de redactar el estado del arte relacionado con el Ensemble.
- Paper "MReC-DFS" -> Naive Bayes as a base learner -> Posible paper a introducir dentro de las propuestas de redes bayesianas.
- Una vez leídos bastantes papers -> Buscar menciones de papers dentro de otros papers (por fecha).

- A la hora de implementar un algoritmo -> SIMULAR FLUJO DE DATOS PARTIENDO LOS DATOS EN BLOQUES DE TAL FORMA QUE SIMULAMOS QUE LOS DATOS VAN LLEGANDO A MEDIDA QUE PASA EL TIEMPO.
- Añadir documento “Online Machine Learning in Big Data Streams (2018)” al grupo de surveys a consultar.
- Pedir libro
- ¿Las referencias que no sean de comparativas hace falta que se ordenen por fecha?
- Ir mandando lo que vaya redactando del TFM.
- En la teoría de las redes bayesianas, ¿tengo que poner mucha notación? ¿O no hace falta?

## REUNIÓN 26 de abril 2019

- Revisar la bibliografía introducida en la memoria (aunque esté en formato bibtex, puede que los títulos no estén bien escritos, con las mayúsculas pertinentes.
- Ver documentos de Javier Diaz para obtener información sobre clustering para data streams. -> Clustering of Data Streams with Dynamic Gaussian Mixture Models. An IoT Application in Industrial Processes
- A la hora de realizar la comparación entre propuestas, centrarnos en las diferencias en cuanto a metodología más que en los experimentos acerca de si un método es mejor que otro.
- Al haber una revisión de ensemble de 2017, no centrarnos en las propuestas que ya están en esta revisión, sino más en aquellas que no están (comparar las propuestas que no están entre ellas y con otras propuestas que están en los métodos de ensemble).
- Añadir en cada una de las comparaciones entre propuestas que tratan un mismo algoritmo de aprendizaje automático para flujos de datos una tabla resumen como las que se encuentra en alguna de las revisiones (por ejemplo si manejan el concept drift).
- Poner poco de teoría de cada uno de los algoritmos de aprendizaje automático.
- Con respecto a las propuestas que traten el aprendizaje semi-supervisado, al ser pocas, meterlas dentro de las demás propuestas.
- Tener en cuenta el orden en el que nombramos las diferentes propuestas. Por ejemplo, sería conveniente poner antes la inducción de reglas que los árboles de decisión.
- Preguntar lo de las referencias relacionadas con las páginas web y las referencias en las imágenes.
- Mencionar lo del entrenamiento en las redes neuronales sobre que me he referido al descenso del gradiente normal, procesando todos los ejemplos primero y luego haciendo backpropagation.

- Mencionar lo de la notación -> En la tabla notación general y según el caso una notación específica (ej: KNN).
- A la hora de hablar de los clasificadores bayesianos, no se si es necesario meter una definición formal de las redes bayesianas. Por otra parte, cuando vaya a hablar de las redes bayesianas para el descubrimiento de conocimiento, no se si debería meter la definición formal ahí.
- Revisar la parte de las fórmulas y de cómo se relacionan en los clasificadores bayesianos.

Thus,

$$p(\mathbf{x}, c) = p(c|\mathbf{pa}(c)) \prod_{i=1}^{n} p(x_i|\mathbf{pa}(x_i)). \quad (2)$$

When the sets $\mathbf{Pa}(X_i)$ are sparse, this factorization prevents having to estimate an exponential number of parameters, which would otherwise be required.

For the special case of $\mathbf{Pa}(C) = \emptyset$, the problem is to maximize on $c$:

$$p(\mathbf{x}, c) = p(c)p(\mathbf{x}|c).$$

Therefore, the different Bayesian network classifiers explained later correspond with different factorizations of $p(\mathbf{x}|c)$. The simplest model is the naive Bayes, where $C$ is the parent of all predictor variables and there are no dependence relationships among them (Sections 3 and 4). We can progressively increase the level of dependence in these relationships (one-dependence, $k$-dependence, etc.) giving rise to a family of augmented naive Bayes models, explained in Sections 5 through 8.1; see Figure 1.

### 6.1. Tree-Augmented Naive Bayes

Unlike in seminaive Bayes, which introduces new features to relax the conditional independence assumption of naive Bayes, the *tree-augmented network* (TAN) [Friedman et al. 1997] maintains the original predictor variables and models relationships of at most order 1 among the variables. Specifically, a tree-shaped graph models the predictor subgraph (Figure 6).

Learning a TAN structure first involves constructing an undirected tree. Kruskal's algorithm [Kruskal 1956] is used to calculate the maximum weighted spanning tree (MWST), containing $n - 1$ edges, where the weight of an edge $X_i - X_j$ is $I(X_i, X_j|C)$, which is the conditional mutual information of $X_i$ and $X_j$ given $C$. The undirected tree is then converted into a directed tree by selecting at random a variable as the root node and replacing the edges by arcs. This is the tree shaping the predictor subgraph. Finally, a naive Bayes structure is superimposed to form the TAN structure. The posterior distribution in Equation (1) is then

$$p(c|\mathbf{x}) \propto p(c)p(x_r|c) \prod_{i=1, i\neq r}^{n} p(x_i|c, x_{j(i)}), \quad (11)$$

La ecuación p(**x**,c)=p(c)p(**x**|c) no se de donde viene. Además, ¿Por qué los diferentes clasificadores bayesianos corresponden con diferentes factorizaciones de p(**x**|c) si la del TAN incluye una variable predictora en la parte condicional (p(x_i|c,x_j(i)))?

- Si las reglas son más generales que los árboles de decisión, ¿deberíamos ponerlas en la teoría después de los árboles de decisión?
- Preguntar si en la teoría de regresión logística es necesario meter lo del logit.
- Mirar tesis de Hanen Borchani para definir mejor el problema de clasificación supervisada. -> Preguntar si es necesario definirlo mejor puesto que al nombrar

la notación no sabemos si hay que mencionar lo que es cada cosa al establecer el problema de clasificación supervisada.

- Paper “Hierarchical Clustering of Time-Series Data Streams” -> ¿En clustering o en time-series?
- ¿Las propuestas para series temporales tienen que tratar con las restricciones de los data streams?
- Comprobar si hay alguna propuesta que aborda algún tipo de algoritmo que no se ha explicado en la teoría.
- ¿Tanto el clustering jerárquico aglomerativo como divisivo producen el mismo dendograma solo que en direcciones opuestas? ¿O pueden construir diferentes dendogramas? -> Saberlo por si hay que cambiar la imagen
- Diferencia entre model-based clsutering y density-based clustering.
- El método de estimación de máxima verosimilitud tiene como objetivo maximizar la probabilidad de pertenencia de los distintos objetos a cada uno de los clusters.
- Paper “McDiarmid Drift Detection Methods for Evolving Data Streams” -> Menciona que el virtual drift se produce cuando hay cambios en la probabilidad P(X) (supongo que será la marginal, no la a priori) y, por tanto, cambios en la probabilidad P(X|C). ¿Por qué ocurre esto?
- Si el virtual drift no afecta a los límites de decisión, ¿por qué se tiene en cuenta?
- Decir que he encontrado un par de surveys mas de clustering (2018), pero que voy a justificarlo diciendo que me voy a centrar en ciertos métodos de clustering, que no abordan métodos de clustering tan recientes y que vamos a proponer más recientes, además de añadir otros que no se encuentran ahí.
- An Evaluation of Data Stream Clustering Algorithms (2018) -> Contiene una tabla comparativa, pero mezclando propuestas de distintos tipos de clustering. Tiene pocos de partinioning. Vamos a proponer hacer tablas distintas comparando diferentes tipos de métodos para clustering.
- Landmark window model -> ¿Se puede modificar el landmark de forma dinámica?
- Hay veces que se menciona que los algoritmos incrementales son también online. Los modelos de inducción de árboles incrementales ID4, ID5, ID5R, etcétera, ¿se pueden considerar que son apropiados para flujos de datos? ¿Los incorporo en la tabla?
- ¿Es necesario meter la definición del Hoeffding Bound en el VFDT? Se pregunta porque si no se va a hacer muy largo lo dedicado a esa propuesta.
- Preguntar si es necesario seguir un orden cronológico en los artículos mencionados.
- OLIN -> The connectionist nature of the info-fuzzy network (each terminal node is connected to every target node) resembles the topological structure of multi-layer neural networks (see [27]), which also have input and output nodes and a variable number of hidden layers. Consequently, we define our model as a network and not as a tree. -> No obstante, lo compara con el CVFDT -> ¿Lo pongo en las propuestas de árboles de decisión?

- UFFT -> Para un problema multiclase construye un bosque de árboles de decisión binario, uno para cada par de valores que puede tomar la variable clase. ¿Es cierto que no es ensemble?
- Los clasificadores Naive Bayes que se utilizan para splitting-tests se sitúan en inner nodes, pero en el paper solo habla de las hojas. -> ¿Se utilizan solo para las hojas?
- En los cuadros que resumen las diferentes propuestas hay muchos huecos vacíos. Mencionarlo.
- RGNBC -> En función de que *concept drift* se produzca, se tienen en cuenta unos atributos u otros y se añaden a la tabla de información. ¿Se puede considerar que manejan la aparición de nuevos atributos?
- Preguntar si el tema de las series temporales y las redes bayesianas para el descubrimiento del conocimiento se puede poner en el trabajo futuro.
- En el KNN, ¿se puede considerar que en todas las propuestas se adaptan a nuevas clases debido a la forma en la que se lleva la clasificación de nuevas instancias? Con respecto al tratamiento de instancias de alta dimensión, ¿se puede poner en la tabla que no se trata por defecto a no ser que se aplique algún método para abordarlas? Debido al problema del cálculo de las distancias en datos de alta dimensión.
- ¿En el KNN es necesario poner en la tabla si se detecta concept drift o no?
- En la propuesta ANNCAD comentan la construcción de varios clasificadores basados en el algoritmo ANNCAD, pero dentro de la sección del KNN no lo comento puesto que lo quiero mencionar en las propuestas de ensemble. ¿Es adecuado hacerlo de esta manera?
- SVM tiene buen desempeño con datos de altas dimensiones. ¿Se puede poner en el cuadro que todas las propuestas manejan datos de altas dimensiones?
- Regresión logística -> Hay un paper que no entiendo bien. ¿Lo pongo?