Active ML

DSLAB

Máster en Data Science. ETSII.

Móstoles, Madrid

June 3, 2022

Universidad
Rey Juan Carlos

Índice

Universidad Rey Juan Carlos

Introducción

Contexto - I

Universidad Rey Juan Carlos

- Los modelos de aprendizaje automático supervisado se entrenan en un conjunto de datos previamente categorizados (etiquetados).
- El proceso de etiquetado es generalmente costoso y limita la cantidad de datos de entrenamiento.
- La recolección/obtención de observaciones (sin etiquetas) también puede ser un proceso costoso.
- Por último, el **tamaño** y la **calidad** del conjunto de datos generalmente limita el rendimiento de los modelos de aprendizaje automático.

Contexto - II

Universidad Rey Juan Carlos

¿Son todas las observaciones igual de importantes en el proceso de creación de un modelo de aprendizaje automático?

Active ML

Universidad Rey Juan Carlos

El **aprendizaje activo** es un caso especial de aprendizaje automático en el que un algoritmo de aprendizaje puede consultar de forma interactiva a un usuario (o alguna otra fuente de información) para **etiquetar nuevos datos**.

Active ML

Universidad Rey Juan Carlos

Técnicamente, tenemos un conjunto de datos etiquetados L y de datos no etiquetados U, y el objetivo es seleccionar un subconjunto de datos a etiquetar S, $S \subseteq U$, de tal forma que entrenar el modelo en el conjunto $L \cup S$ maximice una métrica.

- El número de observaciones que podemos etiquetar se conoce generalmente como presupuesto.
- La técnica a utilizar para seleccionar C dependerá diferentes aspectos como el modelo de aprendizaje automático subyacente, métrica de rendimiento, etiquetado secuencial o por lotes..
- La métrica puede ser de rendimiento, diversidad..
- Oráculo: Entidad que proporciona las etiquetas (personas, otros sistemas...)

Escenarios de aplicación de Active ML

Universidad Rey Juan Carlos

- Colecciones de datos. En internet tenemos grandes cantidades heterogéneas de datos que se pueden utilizar para entrenar modelos en diferentes tareas. Por ejemplo, el audio de los vídeos se podría utilizar para entrenar un sistema de transcripción.
- Privacidad. En determinados escenarios etiquetar datos requiere obtener permiso de una entidad/persona, lo que suele ser un proceso costoso y lento.
- Redes sociales. Previo a una campaña de publicidad en redes sociales se realizan estudios para identificar la población objetivo. Estos estudios involucran encuestas y análisis manuales de publicaciones, que suelen ser procesos muy costosos.

¿Cómo seleccionamos el conjunto de datos?

Universidad Rey Juan Carlos

- Estrategia pasiva. Escoger una o varias observaciones de forma aleatoria.
- Estrategia activa. Utilizar un procedimiento que seleccione las observaciones en base a un criterio.

Escenarios de etiquetado

Universidad Rey Juan Carlos

- *Selective or Sequential Sampling*. Muestreamos sobre la distribución subyacente, y se decide si esas observaciones han de ser etiquetadas o no.
- *Membership Query Synthesis*. Se muestrean observaciones de una distribución que no tiene que coincidir con la distribución subyacente de los datos.
- *Pool-based Sampling*. Seleccionar las instancias mas representativas para el proceso de aprendizaje de un conjunto dado.

Métodos de selección

Uncertainty sampling: Intuición

Universidad Rey Juan Carlos

Intuición: Se etiqueta aquella instancia en la que el modelo está menos "seguro". En un modelo probabilístico binario, ésta observación, x^*, se podría calcular como:

$$x^* = argmin_{x \in U} \ |0.5 - \hat{P}(Y = 1|X = x)|$$

Donde $\hat{P}(Y = 1|X = x)$ es la probabilidad estimada de ser clasificado como 1 dado x. En el caso de una SVM, se podría calcular como la observación cuyo valor en la función de decisión es mas cercano a 0:

$$x^* = argmin_{x \in U} \ |f(x)|$$

Donde f es la función de decisión de la SVM (ver escalado de Platt).

Uncertainty sampling: Multi-clase I

Universidad Rey Juan Carlos

La generalización a multi-clase de este método es trivial (least-confident):

$$x^* = argmax_{x \in U} \ \hat{P}(Y = y^* | X = x)$$

Donde $y^* = argmax_{y \in Y} \hat{P}(Y = y | X = x)$ es la etiqueta mas probable. Sin embargo, ésta generalización solo tiene en cuenta información de la etiqueta mas probable, ignorando el resto. Por ejemplo, en los siguientes dos casos se obtendría el mismo "valor":

- $\hat{P}(Y = 1 | X = x_1) = 0.5$, $\hat{P}(Y = 2 | X = x_1) = 0.49$, $\hat{P}(Y = 3 | X = x_1) = 0.01$
- $\hat{P}(Y = 1 | X = x_2) = 0.5$, $\hat{P}(Y = 2 | X = x_2) = 0.25$, $\hat{P}(Y = 3 | X = x_2) = 0.25$

Pero en el caso 1) la incertidumbre es mayor ya que la diferencia entre las dos clases mayoritarias es mucho menor que en 2).

Uncertainty sampling: Multi-clase II

Universidad Rey Juan Carlos

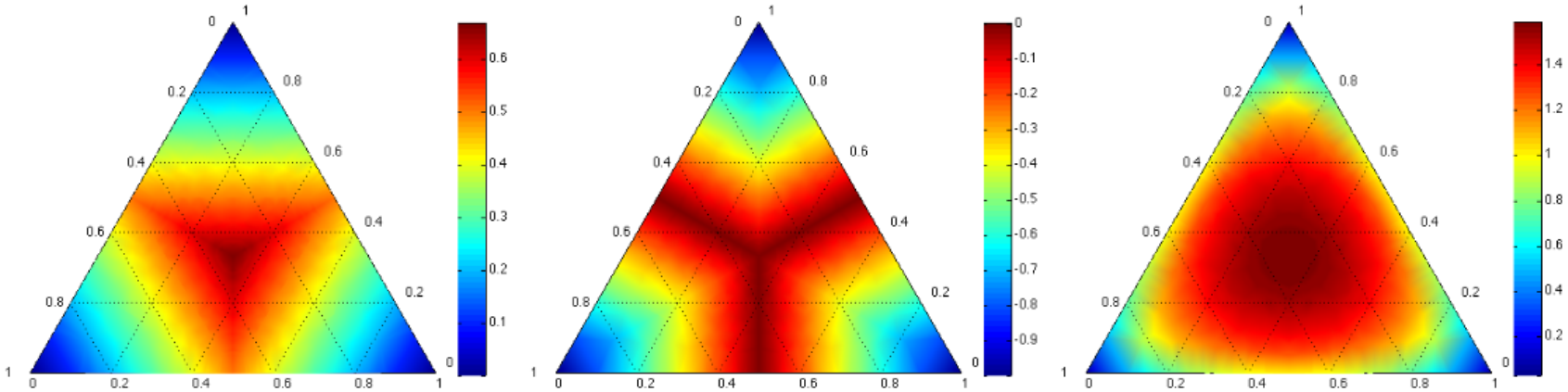
La limitación del método anterior se puede mitigar incorporando información de la segunda etiqueta mas probable (margin):

$$x^* = argmax_{x \in U} \ \hat{P}(Y = y^*|X = x) - \hat{P}(Y = y_2^*|X = x)$$

Donde $y_2^* = argmax_{y \in Y \setminus \{y^*\}} \hat{P}(Y = y|X = x)$ es la segunda etiqueta mas probable. Otra posible técnica es usar la médida de entropía de Shannon (la más popular):

$$x^* = argmax_{x \in U} \ \sum_{y \in Y} \hat{P}(Y = y|X = x) \cdot log(\hat{P}(Y = y|X = x))$$

Uncertainty sampling: Multi-clase III



(a) least confident (b) margin (c) entropy

Ilustración: Comparación de medidas de muestreo por incertidumbre.

Query-by-comittee: Intuición

Universidad Rey Juan Carlos

Intuición: Tenemos una serie de hipótesis diferentes (modelos), y seleccionamos aquella muestra en la cual haya mas disparidad en su etiquetado (regiones de desacuerdo). Con esta selección conseguimos reducir el espacio de hipótesis (modelos compatibles con el actual conjunto de datos).

Query-by-comittee: Implementación

Universidad Rey Juan Carlos

Dependiendo de tipo de modelo, existen varias formas de generar diferentes hipótesis:

- Modelos generativos: Muestrear un conjunto de modelos compatibles de la distribución a posteriori $P(\theta|L \cup S)$.
- Modelos discriminativos: El más común es bagging.

Para medir el desacuardo también existen varias técnicas:

- Vote entropy: $x^* = argmax_{x \in U} \; -\sum y \in Y \frac{V(y|x)}{C} \cdot log(\frac{V(y|x)}{C})$
- KL divergence average: $x^* = argmax_{x \in U} \; C^{-1} \cdot \sum_{i=1}^{C} D(\hat{P}_c || \hat{P})$

Donde C es el número de hipótesis, $V(y|x)$ es el número de votos de la etiqueta y para la observación x, $\hat{P}_c$ es la hipótesis $c \in \{1, \ldots, C\}$, y $\hat{P}$ es el consenso.

Expected model change: Intuición

Universidad Rey Juan Carlos

Intuición: Seleccionar la instancia, que en caso de ser añadida al conjunto de entrenamiento, cambiase más el modelo. ¿Cómo podemos saber cómo impacta una instancia en el entrenamiento sin saber su etiqueta?. La calculamos en base a cómo cambiaría con cada uno de los posibles etiquetados (p.ej., utilizando una media).

Expected model change: Implementación

Universidad Rey Juan Carlos

Un ejemplo es el cambio de magnitud esperada en el gradiente (expected gradient length) en modelos que se entrenan con métodos basados en el gradiente (p.ej., redes neuronales).

$$x^* = argmax_{x \in C} \sum_{y \in Y} \hat{P}(Y = y|X = x) \cdot ||\nabla l_{\hat{P}}(L \cup < x, y >)||$$

Donde $|| \cdot ||$ es la norma euclídea y $l_{\hat{P}}$ es la función de pérdida para el modelo $\hat{P}$.

Expected error reduction: Intuición

Universidad Rey Juan Carlos

Intuición: Es un método similar al anterior, pero en este caso se selecciona la instancia que una vez añadida al conjunto de entrenamiento, minimizaría una métrica de error sobre un conjunto de test. Como el método anterior, dado que no conocemos la etiqueta real, calculamos el valor esperado sobre todas las etiquetas.

Expected error reduction: Implementación

Universidad Rey Juan Carlos

En general, este método es el que "mejor funciona", pero es el método mas costoso computacionalmente, ya que es necesario realizar un re-entrenamiento por cada observación y etiqueta.

$$x^* = argmin_{x \in C} \sum_{y \in Y} \hat{P}(Y = y|X = x) \cdot (\sum_{z \in T} 1 - \hat{P}_{<z,y(z)>}(y(z)|z))$$

Donde T es un conjunto de test, $y(z)$ es la etiqueta para la instancia z, y $\hat{P}_{<z,y(z)>}$ es el modelo después de haber re-entrenado con la instancia z.

Density-weighted methods: Intuición

Universidad Rey Juan Carlos

Intuición: La idea general es que las instancias tienen que provenir de zonas inciertas pero también de zonas representativas de la distribución subyancente. Para tener en cuenta las zonas representativas se utiliza una similitud, penalizando aquellas observaciones que son "poco probables".

Density-weighted methods: Implementación

Universidad Rey Juan Carlos

Una posible implementación de ésta técnica es combinar una estrategia base (como las vistas anteriormente quitando el argmax o argmin) y una similitud (p.ej., similitud coseno o euclidea):

$$x^* = argmax_{x \in U} \phi(x) \cdot ((\#U)^- 1 \sum_{z \in U/\{x\}} sim(z, c))$$

Donde ϕ es una estrategia base, $\#U$ es el cardinal de U, y sim una medida de similitud.

Detalles prácticos

Lotes o secuencial

Universidad Rey Juan Carlos

Generalmente los etiquetados se realizan de forma secuencial, es decir, se selecciona una observación, se etiqueta, se vuelve a seleccionar una observación... En este contexto, un método por lotes es capaz de proporcionar varias instancias para ser etiquetadas a la vez.

- ¿Qué ventajas plantea un método por lotes?
- ¿Se podrían utilizar las técnicas anteriores en un contexto por lotes?

Coste de etiquetado variable

Universidad Rey Juan Carlos

El objetivo principal en el aprendizaje activo es reducir el coste, tanto monetario como temporal, en el que se incurre al etiquetar datos. ¿Qué pasa cuando el coste de etiquetar los datos no es uniforme?.

- Asumir que todos cuestan lo mismo.
- Añadir una penalización que depende del coste de etiquetado. El coste de etiquetado puede o no ser conocido de antemano, en este último caso se puede estimar con un modelo de regresión.

Deriva conceptual

Universidad Rey Juan Carlos

Hasta ahora hemos asumido que el proceso que estamos modelando en el tiempo es estático e inmutable, pero esto raramente es cierto.

- ¿Qué problema plantean las técnicas previas en un contexto cambiante?
- ¿Cómo podemos solucionar este problema?

¿Cuándo parar?

Universidad Rey Juan Carlos

Por último, ¿Tenemos que parar de etiquetar en algún momento?

- Etiquetar nuevas instancias tiene un coste, y nuestro objetivo es reducir el coste total.
- Un posible criterio de parada es cuando el coste de etiquetar es superior al coste de los fallos.
- Por otra parte, en un entorno de streaming no "podemos parar" ya que los posibles cambios de distribución harían que las estimaciones de costes de los fallos fuesen erróneas.

Librerías

Universidad Rey Juan Carlos

- Python: https://github.com/modAL-python/modAL
- R: https://github.com/ramhiser/activelearning

Referencias

Referencias

Universidad Rey Juan Carlos

- https://burrsettles.com/pub/settles.activelearning.pdf
- https://github.com/modAL-python/modAL
- https://www.tsc.uc3m.es/~miguel/MLG/adjuntos/ActiveLearningFinal.pdf
- https://arxiv.org/pdf/2101.11665.pdf