

DIPLOMATURA EN CIENCIA DE DATOS

Valoración de modelos

Valoración de modelos

Temario a tratar:

- ¿Por qué?
- Cuando se realiza
- Tipos de problemas
 - Aprendizaje supervisado
 - Clasificadores
 - Regresores
 - Aprendizaje no supervisado
- Aproximación teórica
 - La matriz de confusión
 - Indicadores adimensionales
- Aproximación práctica
 - La matriz de valoración
 - Valor del modelo, piso y techo
- Conclusiones

¿Por qué?

Cuando enfrentamos la necesidad de crear un modelo no sabemos, a priori, cuál será el algoritmo que nos permita construir el mejor modelo.

La solución pasa por probar con varios algoritmos para ver cual anda mejor. Sin embargo esto de 'anda mejor' es algo que necesitamos definir bien y no resulta para nada evidente.

El mejor algoritmo será el que:

- Termina más rápido
- Ocupa menos memoria
- Tiene una mayor precisión
- Etc...

Se puede discutir interminablemente hasta que recordamos que la ciencia de datos es una ciencia práctica. Resolvemos problemas para organizaciones que tienen un propósito bien definido:

Hospital: curar personas

Banco: ganar dinero

Universidad: enseñar

Entonces, la contribución de nuestro modelo a la realidad del cliente dependerá del impacto que tenga en el objetivo que dicho cliente persigue. Esto nos permitirá escapar de las valoraciones abstractas que, como veremos más abajo, pueden llevarnos a muy malas decisiones.

Cuando se realiza

La valorización de un modelo se realiza:

- Para la sintonía de los hiper-parámetros
- Para seleccionar el modelo que mejor resuelve un problema tras haber sintonizado los hiper-parámetros de cada uno.

Cuando se entrena, por ejemplo, un árbol de decisión, se obtienen distintos resultados dependiendo del parámetro de complejidad, de la cantidad de casos mínimos por hoja, de la profundidad máxima del árbol, de la selección del criterio de división, etc. Estos valores que hay que ajustar para obtener un resultado cambian con cada algoritmo. El nombre genérico que le damos son hiper-parámetros.

Tipos de problemas

Como tenemos distintos modelos vamos a tener que resolver como valorizamos cada caso. En principio podemos considerar:

- Aprendizaje supervisado
 - o Clasificadores
 - o Regresores
- Aprendizaje no supervisado

En el caso del aprendizaje supervisado conocemos un universo con todos sus atributos. Además, para una muestra dentro de ese universo conocemos un resultado. Lo que nos proponemos es predecir ese resultado para el resto del universo por fuera de la muestra apoyándonos en los atributos.

Un clasificador es un modelo con una salida categórica. Para medir el éxito de un clasificador habrá que tener en cuenta los aciertos y los errores de clasificación. Acá asoma la oreja el problema de la valoración. Tratemos de verlo con un ejemplo.

Consideremos lo que pasa en la guardia de un hospital:

- Contamos con recursos limitados para atender a los pacientes que van llegando.
- Algunos tienen problemas que necesitan tratamiento urgente y otros no.
- La primera decisión que la guardia debe tomar es si necesita tratarlo urgentemente o puede esperar. (Esa decisión se conoce como triage)

Se pueden dar cuatro circunstancias:

- Trato en forma urgente un caso que era realmente urgente
- Trato en forma urgente un caso que no era urgente
- Hago esperar a un caso que no era urgente
- Hago esperar a un caso que era realmente urgente

Si nuestro modelo es responsable de predecir la urgencia a cada una de estas circunstancias las llamamos:

Verdadero Positivo = TP = Predigo urgencia y es urgencia

Falso Positivo = FP = Predigo urgencia y no es urgencia

Verdadero Negativo = TN = Predigo no urgencia y no es urgencia

Falso Negativo = FN = Predigo no urgencia y es urgencia

TP y TN son aciertos. Ahí no hay problema, cuantos más aciertos mejor.

FP y FN son errores. El problema es que desde la perspectiva del objetivo del hospital los dos errores no importan lo mismo.

Por el primer error (FP), lo que puede pasar, es que sature temporalmente la capacidad de la guardia. No van a salir en el diario por eso.

Por el segundo error (FN) la cosa cambia dramáticamente, que se muera un paciente urgente en el pasillo de la guardia por un error de triage es dramático.

Si tratamos de optimizar la precisión como indicador abstracto $(TP+TN)/(TP+TN+FP+FN)$ no vamos a tener la posibilidad de preferir cometer un FP a un FN ya que ambos cuentan del mismo modo para ese indicador. Ya vamos a ver cómo podemos superar esa dificultad.

Un regresor es un modelo con una salida numérica. Para medir el valor aportado por un regresor se suele recurrir al RMS (root mean square o, en castellano, raíz del error cuadrático medio)

De acuerdo a eso, al RMS le da lo mismo errar por exceso que por defecto. Hay que notar que para muchos problemas, simplemente no da lo mismo. Vamos a tener que buscar alternativas para esto.

Aproximación teórica

Para volcar el resultado de un clasificador de dos categorías se utiliza la matriz de confusión:

	Predigo Rojo	Predigo Negro
Sale Rojo	TP	FN
Sale Negro	FP	TN

A partir de esta matriz de confusión se pueden definir diversos indicadores:

- Sensibilidad: $TP/(TP+FN)$
- Especificidad: $TN/(TN+FP)$
- VPP: $TP/(FP+VP)$
- VPN: $TP/(TN+FN)$
- Exactitud: $(TP+TN)/(TP+TN+FP+FN)$

Estos indicadores son conocidos en algunos ámbitos y es seguro utilizarlos con los que están imbuidos de su significado. Por ejemplo en el mundo médico resulta bastante natural hablar de que un determinado análisis es sensible o específico a la hora de llegar a un diagnóstico.

Sin embargo debemos tener cuidado de usarlos fuera del contexto en el cual son entendidos porque su carácter abstracto permite interpretaciones que pueden resultar nocivas.

Aproximación práctica

Como mencionamos más arriba las distintas circunstancias que pueden obtenerse como consecuencia de la aplicación de un modelo predictivo pueden tener distinto efecto en los objetivos que persigue la organización con la que estamos colaborando.

Consideremos por un momento que se trata de ayudar a un call center a identificar cuáles son los prospectos con los cuales tenemos mayor probabilidad de venderles.

Simplificando mucho el problema se pueden dar cuatro circunstancias:

	Predigo Venta	Predigo No venta
Sale Venta	VTP = \$ 1000	VFN = \$0
Sale No Venta	VFP = \$ -100	VTN = \$ 0

Esto es lo que llamamos la matriz de valoración.

Los valores que definen la matriz no salen de las cuentas que podamos hacer sobre nuestros datos sino que debemos sacarlos del mismo negocio. Necesitamos que exista una comprensión y un consenso general sobre estos números ya que nuestra solución se volverá totalmente dependiente de ellos.

Valor del modelo, piso y techo

Si multiplicamos la matriz de confusión por la matriz de valoración posición por posición y sumamos el resultado obtenemos el valor que aporta nuestro modelo:

$$V = VTP*TP + VTN*TN + VFP * FP + VFN * FN$$

¿Qué valor de V es bueno?

En principio un número sólo no es ni bueno ni malo ni grande ni chico hasta que no se lo compara con otro.

Nosotros vamos a definir dos puntos de comparación.

El primero es el piso. Representa el estado actual de la solución que ya tiene la organización para ese problema.

Si se trata de dar créditos de acuerdo a un protocolo y a la opinión de un oficial de créditos ese será el punto de comparación.

Si se trata de llamar a los prospectos para venderles y la organización no cuenta con un mecanismo de discriminación de a que teléfonos llamar entonces el piso lo obtendremos llamando a todos.

En general el piso corresponderá a la mejor solución disponible antes de nuestro análisis:

- Llamar a todos.
- No llamar a nadie.

¿Nuestro V puede dar por debajo del piso?

Tristemente sí. En ese caso significa que necesitamos mejorar nuestro modelo, ya sea ajustando los hiper-parámetros o cambiando de algoritmo. Con un valor de V por debajo del piso sería mejor no mostrar nuestro modelo a nadie.

¿Además del piso podemos considerar un techo?

Sí, y además es muy fácil calcular su valor. Basta con considerar que el modelo perfecto sólo puede tener aciertos:

$$V_{techo} = VTP * \text{ventas} + VTN * (\text{no ventas})$$



Esta filosofía se puede extender a clasificadores que tengan más de dos categorías. Tiene una complejidad adicional si los elementos de la matriz de valoración no tienen las mismas unidades (por ejemplo vidas y dinero)

Conclusiones:

Es posible valorizar los resultados en las unidades que describen el propósito de la organización para la que estamos trabajando.

Para los regresores muchas veces convendrá tener en cuenta por lo menos una función de utilidad para los desvíos positivos y otra para los desvíos negativos como en el caso de la predicción de las ventas.