Trabajo práctico 2: A/B Testing

Ph. D. Saúl Calderón Ramírez
Instituto Tecnológico de Costa Rica,
Escuela de Ingeniería en Computación, Programa de Ciencias de Datos,
PAttern Recongition and MAchine Learning Group (PARMA-Group)

9 de octubre de 2023

Fecha de entrega: Lunes 30 de Octubre
Entrega: Un archivo .zip con el código fuente LaTeX o Lyx, el pdf, y un notebook Jupyter, debidamente documentado, con una función definida por ejercicio. A través del TEC-digital.
Modo de trabajo: Grupos de 3 personas.

Resumen

En el presente trabajo práctico introduce el uso de A/B testing para el desarrollo de sistemas de reconocimiento de patrones.

1. A/B Testing

El *A/B testing* es una práctica de uso extendido en la industria para **asesorar la toma de decisiones en el ámbito técnico basada en datos** (diseño de interfaces, sistemas, modelos de aprendizaje automático). El A/B testing define una **variable independiente** para la cual se estudian distintas variantes: la posición de un botón en una interfaz gráfica, una variante de un modelo de aprendizaje automático, el uso de un motor de base de datos específico, etc., con el objetivo de evaluar de forma controlada el efecto de tal variable independiente en una **variable dependiente**: la satisfacción del usuario, la velocidad para llenar un formulario o la tasa de errores al completar tal formulario, la tasa de aciertos, F1-score, etc., de un modelo de aprendizaje automático.

Luego de aplicar las distintas modificaciones a la variable independiente, las cuales llamaremos **tratamientos,** se realizan N réplicas con idénticas condiciones para los K tratamientos a utilizar. Las N réplicas deben realizarse en **circunstancias totalmente controladas**, donde la única diferencia se da cuando se prueba un distinto tratamiento. Posteriormente, se realiza un análisis estadístico de los resultados, para determinar el **tratamiento ganador**. Es usual que se fije uno o más **tratamientos de control**, que sirvan de referencia para el tratamiento a analizar en la prueba, como se ilustra en la Figura 1.

Este sencillo esquema general de diseño experimental permite:

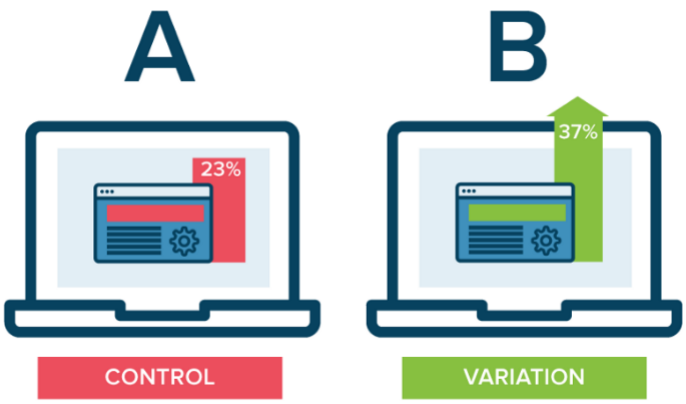

Figura 1: Esquema básico de *A/B testing*.

- **Identificar puntos críticos de mejora en el sistema**: El A/B testing permite introducir cambios pequeños de forma gradual y asesorada, al incentivar la reducción en los términos más simples las variables dependientes que posiblemente afecten en mayor medida a la o las variables independientes.
- **Priorizar modificaciones de bajo riesgo**: La reducción en términos simples de las variables independientes permite generar baterías de pruebas A/B, en las cuales es posible priorizarlas según su riesgo (el cual se puede medir en términos de la estabilidad de la modificación a distintos plazos por ejemplo).
- **Lograr mejoras con significancia estadística, por lo tanto, más fácilmente reproducibles**: El uso de pruebas estadísticas para identificar al tratamiento ganador, mejora la confiabilidad de los resultados y por ende, de la decisión tomada a partir de ellos.

2. A/B Testing para seleccion de caracteristicas

En el equipo de ciencias de datos en el que usted trabaja para resolver el problema de clasificar si un paciente desarrollará diabetes o no, iniciado en el trabajo practico anterior, como continuación de tal trabajo se plantea la siguiente pregunta:

> Es el F1-score mejores al seleccionar las características usando τ de ellas con el puntaje calculado en el trabajo practico anterior?
>
> Dado que el conjunto de datos esta desbalanceado, su equipo se plantea entonces usar como **variable de respuesta** el F1-score

del algoritmo SVC usado en el TP anterior. Para definir el τ a probar, utilice el valor con mejores resultados según el trabajo practico anterior. Para responder tal pregunta, su equipo utilizará el enfoque de *A/B testing*. En tal contexto, se define entonces los siguientes **tratamientos**:

1. El tratamiento A corresponde al uso del algoritmo del perceptrón sin ningún tipo de selección de características. Si en el trabajo practico anterior usó algún preprocesado de datos, utilicelo también.
2. El tratamiento B se refiere al uso del seleccionador de características con los mejores resultados experimentados en el trabajo practico anterior. Respecto al preprocesado, uselo de forma consistente con eltratamiento anterior.

Como **variables de respuesta** se fija el F1-score, los cuales se detallan en el material *Validacion.pdf*.

1. **(5 puntos)** De acuerdo al contexto y pregunta de investigación anteriores, plantee la hipótesis a validar.
2. **(10 puntos)** Implemente en pytorch la métrica del F1-score usando las operaciones básicas de pytorch. Puede comparar lo obtenido con la funcionalidad correspondiente de la librería *scikit learn*.
    a) Documente el diseño y resultados de al menos dos pruebas unitarias de la métrica implementada.

Para desarrollar la comparativa entre ambos métodos, se seguirán los siguientes pasos:

1. **(20 puntos)** Implemente 30 particiones diferentes de los datos, **con reemplazo**, con 70 % de los datos de entrenamiento, y 30 % de los datos de prueba. Asegúrese que las particiones sean la misma utilizada para probar ambos tratamientos.
    a) Documente como se utiliza la herramienta seleccionada para realizar las particiones.
    b) Pruebe ambos tratamientos en las particiones definidas.
    c) Valide que ambos tratamientos utilizan las mismas particiones y muestre la evidencia.
    d) Defina si bajo tales condiciones, los datos están apareados o no (es un *paired-test* o no?).
2. **(10 puntos)** Ejecute las $N = 30$ replicas para los $K = 2$ tratamientos y mida las dos métricas .
    a) Reporte una tabla con los resultados y las siguientes estadísticas descriptivas: **media, mediana, desviación estándar e inclinación**.

b) Comente los resultados y responda las siguientes dos preguntas: Según los estadísticos descriptivos calculados, cree usted que exista un tratamiento con mejora significativa?

c) Además, según tales estadísticos, usted valora posible que las tasas de aciertos de ambos tratamientos sigan una **distribución normal**?

3. **(20 puntos)** Utilizando los resultados obtenidos en el punto anterior, realice las siguientes **gráficas**:

   a) Un histograma de las variables de respuesta por cada tratamiento,

   b) Un diagrama de cajas, con una caja por tratamiento en un mismo gráfico. Ello por cada variable de respuesta.

   c) Un *p-p plot* por cada tratamiento y variable de respuesta.

   d) Comente los resultados y responda las siguientes dos preguntas:

      1) Existe algún valor extremo usando como insumo las visualizaciones anteriores? Explore los datos y defina la razón de ser así.
      2) Según los estadísticos descriptivos calculados, cree usted que exista un tratamiento con mejora significativa?
      3) Además, según tales gráficos, usted valora posible que los puntajes de Dice de ambos tratamientos sigan una distribución normal?

4. **(10 puntos)** Realice una prueba de Kolmogorov-Smirnov y Jarque-Bera por cada tratamiento para verificar si los resultados de cada tratamiento siguen una distribución Gaussiana. Todo ello para los dos variables de respuesta.

   a) Verifique además si hay homocedasticidad usando la función respectiva de *scipy* o alguna librería similar.

5. **(30 puntos)** De acuerdo a las valoraciones de los 3 puntos anteriores, defina si la distribución si las dos variables de respuesta son normales o no.

   a) Si los resultados son normales y presentan homocedasticidad, realice un ANOVA o un t-test pareado para determinar si existe diferencia estadísticamente significativa entrea ambos tratamientos (dependiendo si la prueba califica como *paired test*). Utilice la función *scipy.stats.f_oneway* o *scipy.ttest_rel* de acuerdo a lo anterior. Comente el uso de tal función (entradas y salidas), además de los resultados obtenidos.

   b) Si los resultados no son normales y/o no presentan homocedasticidad, realice un test de Mann-Whitney-U o un Wilcoxon test (dependiendo si usted consideró en el punto anterior si es un *paired-test*) para determinar si existe diferencia estadísticamente significativa entre ambos tratamientos. Utilice la función *scipy.stats.mannwhitneyu* o

scipy.stats.wilcoxon según sea el caso. Comente el uso de tal función (entradas y salidas), además de los resultados obtenidos.

c) Concluya entonces si la el método de selección de características genera una mejora en términos de las dos variables de respuesta. Analice además que otras variables sería importante tomar en cuenta para tomar una decisión final en si utilizar o no la modificación propuesta por Johana.