Bayesian Machine Learning para la investigación científica

Ezequiel Alvarez
sequi@unsam.edu.ar

Éste minicurso está diseñado para dotar a los investigadores en diferentes etapas de su carrera de las herramientas y conocimientos necesarios para aprovechar el Bayesian Machine Learning en su trabajo científico. Al combinar conocimientos teóricos con ejercicios prácticos, el minicurso ofrece una experiencia de aprendizaje equilibrada e integral, diseñada para fomentar tanto la comprensión como la aplicación.

Un conjunto de 5 x (1+2)hs clases que introducirán a los participantes en el mundo del Aprendizaje Automático Bayesiano con fines científicos. El minicurso está adaptado tanto a investigadores senior como junior, atendiendo a sus respectivos niveles de experiencia e interés.

En un bloque teórico de 45' pretendemos transmitir el panorama general del tema de la clase con un enfoque adicional en los detalles y sutilezas desde el punto de vista de la supervisión. Aquí se abordan correctamente los puntos finos y las sofisticaciones, pero sin demostraciones estrictas ni código suministrado. Este bloque está destinado a seniors y juniors, para los seniors como resumen que proporciona los conocimientos para aplicar estas herramientas en la investigación científica; mientras que para los juniors como comprensión y entrada al segundo bloque en el que ponemos las manos en la masa. Concluimos el bloque con una pausa-café prolongada en la que esperamos que las ideas propuestas desencadenen discusiones en torno al campo de estudio de cada participante y cómo aplicarlo en sus datos.

El segundo bloque (~2hs) es muy *hands-on*, está pensado para juniors, pero los seniors interesados en involucrarse activamente en los cálculos también son bienvenidos. Presentamos, discutimos y escribimos código. Los participantes realizan ejercicios de codificación y discuten aplicaciones prácticas. Este bloque hace hincapié en las habilidades prácticas y la resolución de problemas del mundo real. Se utilizan diferentes bibliotecas y programas estadísticos especialmente diseñados para abordar los problemas planteados.
En general, el minicurso está diseñado para cualquier carrera científica. Usamos sobre todo ejemplos de física, pero lo aprendido es útil y perspicaz para cualquier otro campo con investigación científica dura. Intentamos adaptar y discutir los problemas dentro de los campos de investigación de los participantes.

Requisitos previos:
Se espera que los participantes tengan cursos de álgebra y análisis, un buen manejo de vectores y expresiones multidimensionales, algunos conocimientos sobre probabilidad y estadística, y que estén preparados para el razonamiento y el pensamiento abstracto no trivial. Además, se espera que los participantes tengan algún conocimiento de Python o que hayan codificado scripts en cualquier lenguaje.

Programa

Lecture 1: Introducción a técnicas Bayesianas

- **Teoría (45'):** Teorema de Bayes, campos de aplicación (juegos, puzzles, problemas, aprendizaje automático, etc.). Teorema de Bayes en la investigación científica: Aprendizaje automático bayesiano y flujo de trabajo bayesiano. Sustitución de las redes neuronales por técnicas bayesianas cuando las simulaciones no son suficientemente fiables. Aprendizaje a partir de los datos. Modelos gráficos. Modelos de mezclas como problema general al que se enfrentan los científicos. Algoritmos para abordar los modelos de mezclas. Mezclas gaussianas simples
- **Hands-on (~2hs):** Introducción al lenguaje estadístico STAN. Resolución de problemas básicos de inferencia con STAN. Mezcla Gaussiana.

Lecture 2: Ejemplos Bayesianos simples

- **Teoría (45'):** Problemas sencillos bien conocidos de Aprendizaje Automático Bayesiano (Ocho-escuelas, etc). Parámetros, hiperparámetros y Bayes jerárquico. Mezcla Bernoulli. Priors autoconjugadas en un problema de recuento simple. Comprender cómo muchos datos sobrescriben los priors.
- **Hands-on (~2hs):** Notebooks y análisis numéricos para resolver los problemas presentados en la sección de Teoría

Lecture 3: Mixture Models

- **Teoría (45'):** Estructura interna de los datos. Variables latentes. Modelos gráficos. Construcción de funciones de densidad de probabilidad (FDP) no triviales a partir de FDP triviales. Modelos de mezclas. Variables condicionalmente independientes. Expresión explícita de la verosimilitud en los modelos de mezclas. Ejemplos explícitos de modelos de mezclas, por ejemplo pp > hh > bbAA.
- **Hands-on (~2hs):** Introducción a diferentes distribuciones (Dirichlet, exponencial truncada, Normal truncada, etc.) Resolución de problemas reales de modelos de mezcla utilizando STAN. Hacks y trucos en Modelos de Mezcla.

Lecture 4: Consistencia entre la data, el modelo y los resultados

- **Teoría (45')**: Cómo comprobar si los resultados tienen sentido. Comprobación de la modelización con los datos. Probabilidad de los datos. Comprobación predictiva posterior.
- **Hands-on (~2hs)**: Aplicación de la comprobación predictiva posterior. Hacer afirmaciones sobre la modelización y sobre los resultados de la inferencia. Comprobación de la insesgadez en muestreos en cadena. Rhat.

Lecture 5: Mixture Model para distribuciones no paramétricas

- **Theory (45')**: Inclusión de priors estructurados para aprovechar las propiedades esperadas en las distribuciones. Aprovechamiento de la continuidad en las distribuciones. Procesos gaussianos. Explotación de la unimodalidad en las distribuciones. Etiquetado en modelos de mezclas, comparación de curvas ROC. Limitaciones y adaptaciones en los Modelos de Mezcla presentados. Discusión sobre variables no condicionalmente independientes, uso y modelización de correlaciones.

- **Hands-on (~2hs)**: Scripts para inferir sobre priors estructurados. Muestreo de distribuciones suaves y unimodales. Paralelización de la programación de la inferencia para conjuntos de datos complejos en STAN.