Una introducción suave al método Bootstrap

por **Jason Brownlee** el 25 de mayo de 2018 en https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/

El método bootstrap es una técnica de remuestreo que se utiliza para estimar estadísticas sobre una población mediante el muestreo de un conjunto de datos con reemplazo.

Se puede usar para estimar estadísticas de resumen, como la media o la desviación estándar. Se utiliza en el aprendizaje automático aplicado para estimar la habilidad de los modelos de aprendizaje automático al hacer predicciones sobre datos no incluidos en los datos de entrenamiento.

Una propiedad deseable de los resultados de la estimación de la habilidad del modelo de aprendizaje automático es que la habilidad estimada puede presentarse con intervalos de confianza, una característica que no está disponible fácilmente con otros métodos como la validación cruzada.

En este tutorial, descubrirá el método de remuestreo bootstrap para estimar la habilidad de los modelos de aprendizaje automático en datos desconocidos.

Después de completar este tutorial, sabrás que:

- El método bootstrap consiste en remuestrear iterativamente un conjunto de datos con reemplazo.
- Al usar el bootstrap debes elegir el tamaño de la muestra y el número de repeticiones.

Empecemos.

Tutorial general

Este tutorial está dividido en 4 partes; son:

1. Método Bootstrap
2. Configuración de la Bootstrap
3. Ejemplo

Método Bootstrap

El método bootstrap es una técnica estadística para estimar cantidades sobre una población promediando estimaciones múltiples muestras más pequeñas.

Es importante destacar que las muestras se construyen sacando de una en una observaciones de una muestra de datos de gran tamaño y devolviéndolas a la muestra de datos después de haber sido elegidas. Esto permite que una observación dada se incluya en una muestra pequeña dada más de una vez. Este enfoque de muestreo se llama muestreo con reemplazo.

El proceso para construir una muestra se puede resumir como sigue:

1. Elija el tamaño de la muestra.
2. Mientras que el tamaño de la muestra es menor que el tamaño elegido
    1. Selecciona aleatoriamente una observación del conjunto de datos
    2. Añádela a la muestra.

El método bootstrap se puede usar para estimar un parámetro de una población. Esto se hace tomando muestras pequeñas repetidamente, calculando la estadística y tomando el promedio de las estadísticas calculadas. Podemos resumir este procedimiento de la siguiente manera:

1. Elija una serie de muestras bootstrap para realizar
2. Elija un tamaño de muestra
3. Para cada muestra bootstrap.
    1. Sacar una muestra con reemplazo con el tamaño elegido.
    2. Calcula la estadística sobre la muestra.
4. Calcular la media de las estadísticas de la muestras.

El procedimiento también se puede utilizar para estimar la habilidad de un modelo de aprendizaje automático.

> *El bootstrap es una herramienta estadística ampliamente aplicable y extremadamente poderosa que se puede usar para cuantificar la incertidumbre asociada con un estimador dado o un método de aprendizaje estadístico.*

- Página 187, Introducción al aprendizaje estadístico , 2013.

Esto se hace entrenando el modelo con la muestra y evaluando la habilidad del modelo en aquellas muestras no incluidas en la muestra. Estas muestras no incluidas en una muestra dada se denominan muestras fuera de bolsa, o OOB, para abreviar.

Este procedimiento de uso del método bootstrap para estimar la habilidad del modelo se puede resumir de la siguiente manera:

1. Elija una serie de muestras bootstrap para realizar
2. Elija un tamaño de muestra
3. Para cada muestra bootstrap.
    1. Sacar una muestra con reemplazo con el tamaño elegido.
    2. Ajustar un modelo en la muestra de datos.
    3. Calcule la habilidad del modelo con los datos de fuera de bolsa.
4. Calcule la media de las estimaciones de habilidades de los modelos creados con cada muestra

> *Las muestras no seleccionadas generalmente se conocen como muestras "fuera de la bolsa". Para una iteración dada de remuestreo bootstrap, se construye un modelo sobre las muestras seleccionadas y se usan para evaluarlo las muestras fuera de la bolsa.*

- Página 72, Modelización Predictiva Aplicada , 2013.

Es importante destacar que cualquier preparación de datos antes de ajustar el modelo o ajustar el hiperparámetro del modelo debe ocurrir dentro del bucle for en la muestra de datos. Esto es para evitar la fuga de datos cuando se utiliza el conocimiento del conjunto de datos de prueba para mejorar el modelo. Esto, a su vez, puede resultar en una estimación optimista de la habilidad del modelo.

Una característica útil del método bootstrap es que la muestra resultante de estimaciones a menudo forma una distribución gaussiana. Además de resumir esta distribución con una tendencia central, se pueden dar medidas de varianza, como la desviación estándar y el error estándar. Además, un intervalo de confianza se puede calcular y utilizar para acotar la estimación presentada. Esto es útil cuando se presenta la habilidad estimada de un modelo de aprendizaje automático.

Configuración del Bootstrap

Hay dos parámetros que se deben elegir al realizar la rutina de arranque: el tamaño de la muestra y el número de repeticiones del procedimiento a realizar.

Tamaño de la muestra

En el aprendizaje automático, es común usar un tamaño de muestra que sea el mismo que el conjunto de datos original.

> *La muestra de bootstrap es del mismo tamaño que el conjunto de datos original. Como resultado, algunas muestras se representarán varias veces en la muestra de arranque, mientras que otras no se seleccionarán en absoluto.*

- Página 72, Modelización Predictiva Aplicada , 2013.

Si el conjunto de datos es enorme y la eficiencia computacional es un problema, se pueden usar muestras más pequeñas, como el 50% u 80% del tamaño del conjunto de datos.

Repeticiones

El número de repeticiones debe ser lo suficientemente grande para garantizar que se puedan calcular estadísticas significativas, como la media, la desviación estándar y el error estándar en la muestra.

Un mínimo puede ser de 20 o 30 repeticiones. Los valores más pequeños que se pueden usar agregarán más varianza a las estadísticas calculadas en la muestra de valores estimados.

Idealmente, la muestra de estimaciones sería lo más grande posible, dados los recursos de tiempo, con cientos o miles de repeticiones.

Ejemplo

Podemos demostrar el procedimiento de bootstrap con un pequeño ejemplo. Trabajaremos a través de una iteración del procedimiento.

Imagina que tenemos un conjunto de datos con 6 observaciones:

```
1 [0.1, 0.2, 0.3, 0.4, 0.5, 0.6]
```

El primer paso es elegir el tamaño de la muestra. Aquí, vamos a utilizar 4.

A continuación, debemos elegir al azar la primera observación del conjunto de datos. Vamos a elegir 0.2.

```
1 sample = [0.2]
```

Esta observación se devuelve al conjunto de datos y repetimos este paso 3 veces más.

```
1 sample = [0.2, 0.1, 0.2, 0.6]
```

Ahora tenemos nuestra muestra de datos. El ejemplo demuestra a propósito que el mismo valor puede aparecer cero, una o más veces en la muestra. Aquí la observación 0.2 aparece dos veces.

Luego se puede calcular una estimación sobre la muestra dibujada.

```
1 statistic = calculation([0.2, 0.1, 0.2, 0.6])
```

Las observaciones no elegidas para la muestra pueden usarse como observaciones fuera de la muestra.

```
oob = [0.3, 0.4, 0.5]
```

En el caso de evaluar un modelo de aprendizaje automático, el modelo se ajusta a la muestra dibujada y se evalúa a la muestra fuera de la bolsa.

```
train = [0.2, 0.1, 0.2, 0.6]
test = [0.3, 0.4, 0.5]
model = fit(train)
statistic = evaluate(model, test)
```

Con esto concluye una repetición del procedimiento. Se puede repetir 30 o más veces para obtener una muestra de estadísticas calculadas.

```
statistics = [...]
```

Esta muestra de estadísticas se puede resumir calculando una media, una desviación estándar u otros valores de resumen para obtener una estimación final utilizable de la estadística.

```
estimate = mean([...])
```

Extensiones

Esta sección enumera algunas ideas para ampliar el tutorial que tal vez desee explorar.

- Enumere 3 estadísticas de resumen que podría estimar utilizando el método bootstrap.
- Encuentre 3 trabajos de investigación que utilicen el método bootstrap para evaluar el rendimiento de los modelos de aprendizaje automático.
- Implemente su propia función para crear una muestra y una muestra fuera de bolsa con el método bootstrap.

Si exploras alguna de estas extensiones, me encantaría saberlo.

Otras lecturas

Esta sección proporciona más recursos sobre el tema si desea profundizar.

Mensajes

- Cómo calcular los intervalos de confianza de Bootstrap para los resultados del aprendizaje automático en Python

Libros

- Modelado predictivo aplicado , 2013.
- Una introducción al aprendizaje estadístico , 2013.
- Una introducción a Bootstrap , 1994.

API

- sklearn.utils.resample () API
- sklearn.model_selection: API de selección de modelo