Data-Snooping Bias and the Beauty of Linearity

Seminario EMD,TWS, Trading, etc

Septiembre de 2020, Andrés Felipe Téllez

Data-snooping y overfitting 1

- Parece que el autor quiere realmente referirse a overfitting, en vez de data-snooping.
- Overfitting es utilizar muchos más parámetros de los que el modelo requiere, sólo para ajustar de manera mas fiel los datos observados.
- Data-Snooping es el mal uso del análisis de datos para encontrar patrones y correlaciones estadísticamente significativas, incrementando el riesgo de falsos positivos.

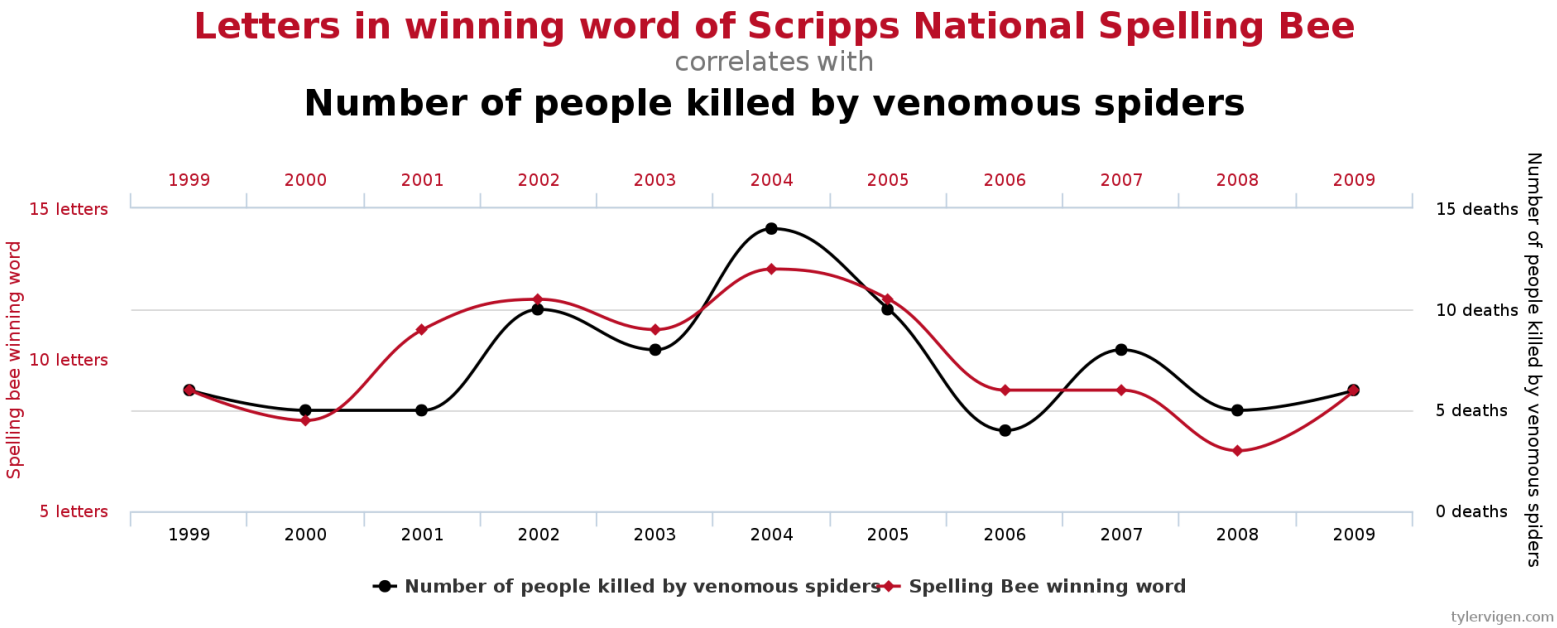
Data-snooping y overfitting 1

- Parece que el autor quiere realmente referirse a overfitting, en vez de data-snooping.
- Overfitting es utilizar muchos más parámetros de los que el modelo requiere, sólo para ajustar de manera mas fiel los datos observados.
- Data-Snooping es el mal uso del análisis de datos para encontrar patrones y correlaciones estadísticamente significativas, incrementando el riesgo de falsos positivos.

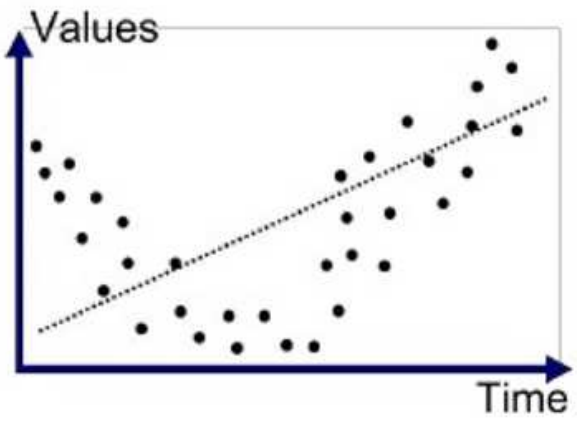
Data-snooping y overfitting 1

- Parece que el autor quiere realmente referirse a overfitting, en vez de data-snooping.
- Overfitting es utilizar muchos más parámetros de los que el modelo requiere, sólo para ajustar de manera mas fiel los datos observados.
- Data-Snooping es el mal uso del análisis de datos para encontrar patrones y correlaciones estadísticamente significativas, incrementando el riesgo de falsos positivos.

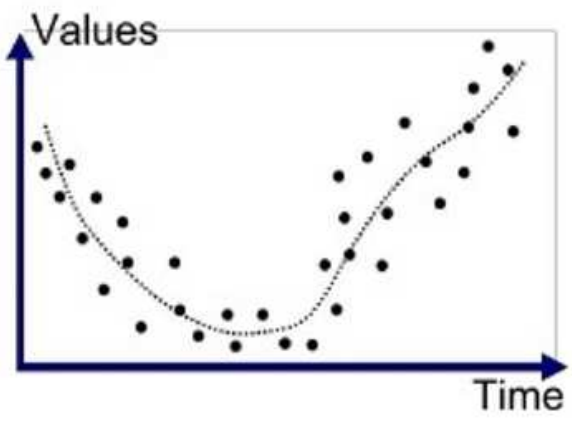
Data-snooping y overfitting 2



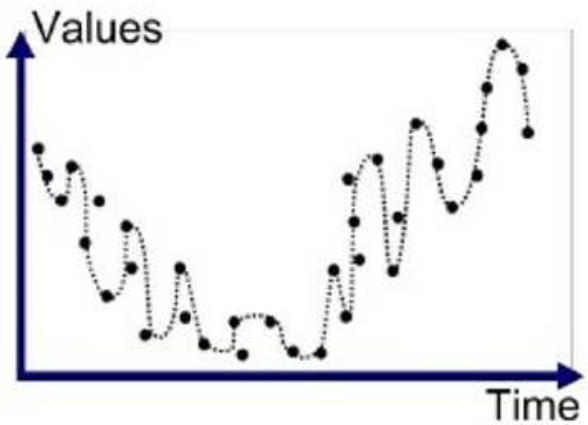
Data-snooping y overfitting 3



Underfitted



Good Fit/Robust



Overfitted

Overfitting

Una herramienta util para evitar el overfitting son los criterios como el AIC y el BIC (nota: en modelos ARMA hay que tener cuidado)

$$AIC = 2k - 2ln(L)$$

$$BIC = kln(n) - 2ln(L)$$

Sobre la linealidad y "Gaussianidad"

- Un modelo lineal no sólo se refiere a una línea recta
- Una parábola también se puede modelar como un sistema lineal
- Un conjunto de ecuaciones diferenciales acopladas también es un modelo lineal
- Los modelos ARMA también.
- Dada su estacionariedad las IMFs (distintas del residuo) son aptas para modelarse con modelos lineales.

Sobre la linealidad y "Gaussianidad"

- Un modelo lineal no sólo se refiere a una línea recta
- Una parábola también se puede modelar como un sistema lineal
- Un conjunto de ecuaciones diferenciales acopladas también es un modelo lineal
- Los modelos ARMA también.
- Dada su estacionariedad las IMFs (distintas del residuo) son aptas para modelarse con modelos lineales.

Sobre la linealidad y "Gaussianidad"

- Un modelo lineal no sólo se refiere a una línea recta
- Una parábola también se puede modelar como un sistema lineal
- Un conjunto de ecuaciones diferenciales acopladas también es un modelo lineal
- Los modelos ARMA también.
- Dada su estacionariedad las IMFs (distintas del residuo) son aptas para modelarse con modelos lineales.

Sobre la linealidad y "Gaussianidad"

- El autor justifica el uso de la linealidad por si simpleza, y para evitar el overfitting, pero se puede caer en el underfitting.
- Hace lo mismo con modelar los retornos de manera Gaussiana, aunque ha sido ampliamente comprobado que no son los mejores para modelar rendimientos.
- Estos dos conceptos (linealidad y distribuciones Gaussianas) si están muy relacionados, y precisamente es uno de los resultados del principio de máxima entropía (o mínima información): Si sólo contamos con la media, la varianza y sabemos que la variable aleatoria toma valores en todos los reales, entonces la distribución gaussiana es la más apta para los rendimientos y además, la estimación lineal es la que minimiza el error cuadrático medio.

Sobre la linealidad y "Gaussianidad"

- El autor justifica el uso de la linealidad por si simpleza, y para evitar el overfitting, pero se puede caer en el underfitting.
- Hace lo mismo con modelar los retornos de manera Gaussiana, aunque ha sido ampliamente comprobado que no son los mejores para modelar rendimientos.
- Estos dos conceptos (linealidad y distribuciones Gaussianas) si están muy relacionados, y precisamente es uno de los resultados del principio de máxima entropía (o mínima información): Si sólo contamos con la media, la varianza y sabemos que la variable aleatoria toma valores en todos los reales, entonces la distribución gaussiana es la más apta para los rendimientos y además, la estimación lineal es la que minimiza el error cuadrático medio.

Sobre la linealidad y "Gaussianidad"

- El autor justifica el uso de la linealidad por si simpleza, y para evitar el overfitting, pero se puede caer en el underfitting.
- Hace lo mismo con modelar los retornos de manera Gaussiana, aunque ha sido ampliamente comprobado que no son los mejores para modelar rendimientos.
- Estos dos conceptos (linealidad y distribuciones Gaussianas) si están muy relacionados, y precisamente es uno de los resultados del principio de máxima entropía (o mínima información): Si sólo contamos con la media, la varianza y sabemos que la variable aleatoria toma valores en todos los reales, entonces la distribución gaussiana es la más apta para los rendimientos y además, la estimación lineal es la que minimiza el error cuadrático medio.

Ejemplo de estrategia lineal

- Contamos con varios factores f_i, cada uno en forma de serie de tiempo, pueden ser: el rendimiento hasta el dia de hoy, el índice de volatilidad, etc
- El objetivo es usarlos para decidir si el rendimiento de mañana R es positivo.
- Primer paso es "estandarizar" los factores (todos en la misma escala):

$$z(i) = \frac{1}{std(f)}(f(i) - mean(f))$$

Luego se hace una "democracia" para predecir R, donde $sign(i)$ es el signo de correlación histórica entre $f(i)$ y R.

$$R = mean(R) + std(R)K$$

$$R = mean(R) + std(R)\left(\frac{1}{n}\sum_{i=1}^{n} sign(i)z(i)\right)$$

Ejemplo de estrategia lineal

- Contamos con varios factores f_i, cada uno en forma de serie de tiempo, pueden ser: el rendimiento hasta el dia de hoy, el índice de volatilidad, etc
- El objetivo es usarlos para decidir si el rendimiento de mañana R es positivo.
- Primer paso es "estandarizar" los factores (todos en la misma escala):

$$z(i) = \frac{1}{std(f)}(f(i) - mean(f))$$

Luego se hace una "democracia" para predecir R, donde $sign(i)$ es el signo de correlación histórica entre $f(i)$ y R.

$$R = mean(R) + std(R)K$$

$$R = mean(R) + std(R)\left(\frac{1}{n}\sum_{i=1}^{n} sign(i)z(i)\right)$$

Ejemplo de estrategia lineal

- Contamos con varios factores f_i, cada uno en forma de serie de tiempo, pueden ser: el rendimiento hasta el dia de hoy, el índice de volatilidad, etc
- El objetivo es usarlos para decidir si el rendimiento de mañana R es positivo.
- Primer paso es "estandarizar" los factores (todos en la misma escala):

$$z(i) = \frac{1}{std(f)}(f(i) - mean(f))$$

Luego se hace una "democracia" para predecir R, donde $sign(i)$ es el signo de correlación histórica entre $f(i)$ y R.

$$R = mean(R) + std(R)K$$

$$R = mean(R) + std(R)\left(\frac{1}{n}\sum_{i=1}^{n} sign(i)z(i)\right)$$

Ejemplo de estrategia lineal

- Contamos con varios factores f_i, cada uno en forma de serie de tiempo, pueden ser: el rendimiento hasta el dia de hoy, el índice de volatilidad, etc
- El objetivo es usarlos para decidir si el rendimiento de mañana R es positivo.
- Primer paso es "estandarizar" los factores (todos en la misma escala):

$$z(i) = \frac{1}{std(f)}(f(i) - mean(f))$$

Luego se hace una "democracia" para predecir R, donde $sign(i)$ es el signo de correlación histórica entre $f(i)$ y R.

$$R = mean(R) + std(R)K$$

$$R = mean(R) + std(R)\left(\frac{1}{n}\sum_{i=1}^{n} sign(i)z(i)\right)$$

Ejemplo de estrategia lineal

- El autor comenta que esta forma de estimación de los rendimientos no siempre certera en términos absolutos para un stock
- Pero sí hace una buena estimación en términos relativos comparando diferentes stocks.
- Es decir, el procedimiento es calcular el rendimiento de cada stock y luego organizarlos del mayor al menor.

Ejemplo de estrategia lineal

- El autor comenta que esta forma de estimación de los rendimientos no siempre certera en términos absolutos para un stock
- Pero sí hace una buena estimación en términos relativos comparando diferentes stocks.
- Es decir, el procedimiento es calcular el rendimiento de cada stock y luego organizarlos del mayor al menor.

Ejemplo de estrategia lineal

- El autor comenta que esta forma de estimación de los rendimientos no siempre certera en términos absolutos para un stock
- Pero sí hace una buena estimación en términos relativos comparando diferentes stocks.
- Es decir, el procedimiento es calcular el rendimiento de cada stock y luego organizarlos del mayor al menor.

Ejemplo de estrategia lineal

- Otro procedimiento que nombra es: organizar los stocks, pero solo con base en un factor, obtenemos su lugar en la lista $rank_s(i)$ para un factor específico $f(i)$. Finalmente, se suman estos rangos con signo (basado en el signo de su correlación con el factor $f(i)$)

$$rank_s = \sum_{i=1}^{n} sign(i)\, rank_s(i)$$

- Un ejemplo de esto es la fórmula mágica de Joel Greenblatt de dos factores:
- $f(1) =$ rendimiento en capital
- $f(2) =$"earnings yield" $=$ ganancias por acción / precio de la acción.

- Se hace el procedimiento anterior y se compran las mejores 30 acciones y se mantienen por un año. Desde 1988 hasta el 2004 ésta estrategia tuvo un APR del 30,8 %, que es significativamente más alto que el rendimiento del $S\&P500$ que fue de 12,4 % en el mismo periodo.

Ejemplo de estrategia lineal

- Otro procedimiento que nombra es: organizar los stocks, pero solo con base en un factor, obtenemos su lugar en la lista $rank_s(i)$ para un factor específico $f(i)$. Finalmente, se suman estos rangos con signo (basado en el signo de su correlación con el factor $f(i)$)

$$rank_s = \sum_{i=1}^{n} sign(i)\, rank_s(i)$$

- Un ejemplo de esto es la fórmula mágica de Joel Greenblatt de dos factores:
- $f(1) =$ rendimiento en capital
- $f(2) =$"earnings yield" = ganancias por acción / precio de la acción.
- Se hace el procedimiento anterior y se compran las mejores 30 acciones y se mantienen por un año. Desde 1988 hasta el 2004 ésta estrategia tuvo un APR del 30,8 %, que es significativamente más alto que el rendimiento del $S\&P500$ que fue de 12,4 % en el mismo periodo.

Ejemplo de estrategia lineal

- Otro procedimiento que nombra es: organizar los stocks, pero solo con base en un factor, obtenemos su lugar en la lista $rank_s(i)$ para un factor específico $f(i)$. Finalmente, se suman estos rangos con signo (basado en el signo de su correlación con el factor $f(i)$)

$$rank_s = \sum_{i=1}^{n} sign(i)\, rank_s(i)$$

- Un ejemplo de esto es la fórmula mágica de Joel Greenblatt de dos factores:
- $f(1) =$ rendimiento en capital
- $f(2) =$"earnings yield" $=$ ganancias por acción / precio de la acción.

- Se hace el procedimiento anterior y se compran las mejores 30 acciones y se mantienen por un año. Desde 1988 hasta el 2004 ésta estrategia tuvo un APR del 30,8 %, que es significativamente más alto que el rendimiento del $S\&P500$ que fue de 12,4 % en el mismo periodo.

¡FIN!