

Ensemble Learning

Descomposición en Sesgo y Varianza

Luis Norberto Zúñiga Morales

28 de febrero de 2023

Evaluación de desempeño de modelos

Desde un punto de vista estadístico, las variables $\mathbf{x}_i \in \mathcal{X}$ y $y_i \in \mathcal{Y}$ son variables aleatorias que toman valores de una distribución de probabilidad $P(X, Y)$.

Evaluación de desempeño de modelos

Desde un punto de vista estadístico, las variables $\mathbf{x}_i \in \mathcal{X}$ y $y_i \in \mathcal{Y}$ son variables aleatorias que toman valores de una distribución de probabilidad $P(X, Y)$.

Pregunta

¿Qué significa $P(X = \mathbf{x}, Y = y)$?

Evaluación de desempeño de modelos

Desde un punto de vista estadístico, las variables $\mathbf{x}_i \in \mathcal{X}$ y $y_i \in \mathcal{Y}$ son variables aleatorias que toman valores de una distribución de probabilidad $P(X, Y)$.

Pregunta

¿Qué significa $P(X = \mathbf{x}, Y = y)$?

Respuesta

La probabilidad de que las variables aleatorias X y Y tomen los valores \mathbf{x} y y si se toma un objeto al azar del universo ω .

Evaluación de desempeño de modelos

Utilizamos un algoritmo \mathcal{A} para aprender un modelo $\varphi_{\mathcal{L}}$ que aprende del conjunto de datos \mathcal{L} , que busca minimizar el valor esperado del error:

Definición: Valor esperado del error de predicción

El valor esperado del error, también llamado error de generalización o error de prueba, del modelo $\varphi_{\mathcal{L}}$ es

$$\text{Err}(\varphi_{\mathcal{L}}) = \mathbb{E}_{X,Y}[L(Y, \varphi_{\mathcal{L}}(X))] \quad (1)$$

donde L es una función de pérdida que mide la discrepancia entre dos argumentos.

Evaluación de desempeño de modelos

Para el caso de clasificación, la función de pérdida más común es la **cero-uno** definida como:

$$L(Y, \varphi_{\mathcal{L}}(X)) = 1(Y \neq \varphi_{\mathcal{L}}(X)) \quad (2)$$

por lo que el error esperado se vuelve la probabilidad de clasificar erróneamente un dato:

$$\text{Err}(\varphi_{\mathcal{L}}) = \mathbb{E}_{X,Y}[1(Y \neq \varphi_{\mathcal{L}}(X))] = P(Y \neq \varphi_{\mathcal{L}}(X)) \quad (3)$$

¿Por qué?: Esperanza de la función indicadora.

Evaluación de desempeño de modelos

Para el caso de regresión, vamos a usar el **error cuadrático**:

$$L(Y, \varphi_{\mathcal{L}}(X)) = (Y - \varphi_{\mathcal{L}}(X))^2 \quad (4)$$

Con esta función de pérdida, el error de generalización se vuelve:

$$\text{Err}(\varphi_{\mathcal{L}}) = \mathbb{E}_{X,Y}[(Y - \varphi_{\mathcal{L}}(X))^2] \quad (5)$$

Evaluación de desempeño de modelos

- En práctica, la distribución de probabilidad $P(X, Y)$ no se conoce, lo que hace que estimar $\text{Err}(\varphi_{\mathcal{L}})$ sea imposible.

Evaluación de desempeño de modelos

- En práctica, la distribución de probabilidad $P(X, Y)$ no se conoce, lo que hace que estimar $\text{Err}(\varphi_{\mathcal{L}})$ sea imposible.
- Tampoco es posible obtener un conjunto de datos virtualmente infinito para estimar empíricamente $\text{Err}(\varphi_{\mathcal{L}})$.

Evaluación de desempeño de modelos

- En práctica, la distribución de probabilidad $P(X, Y)$ no se conoce, lo que hace que estimar $\text{Err}(\varphi_{\mathcal{L}})$ sea imposible.
- Tampoco es posible obtener un conjunto de datos virtualmente infinito para estimar empíricamente $\text{Err}(\varphi_{\mathcal{L}})$.
- ¿Cómo se estima $\text{Err}(\varphi_{\mathcal{L}})$?

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

Vamos a definir $\overline{E}(\varphi_{\mathcal{L}}, \mathcal{L}')$ como el error promedio del modelo $\varphi_{\mathcal{L}}$ sobre el conjunto \mathcal{L}' (puede ser diferente al conjunto \mathcal{L} que se utilizó para entrenar $\varphi_{\mathcal{L}}$) como:

$$\overline{E}(\varphi_{\mathcal{L}}, \mathcal{L}') = \frac{1}{N'} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{L}'} L(y_i, \varphi_{\mathcal{L}}(\mathbf{x}_i)) \quad (6)$$

donde N' es el tamaño del conjunto \mathcal{L}' .

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- La primer (y más sencilla) forma de estimar el error de generalización es la estimación por resustitución o estimación del conjunto de entrenamiento.

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- La primer (y más sencilla) forma de estimar el error de generalización es la estimación por resustitución o estimación del conjunto de entrenamiento.
- Consiste en estimar empíricamente $\text{Err}(\varphi_{\mathcal{L}})$ con el mismo conjunto de datos \mathcal{L} que se utiliza para construir $\varphi_{\mathcal{L}}$:

$$\widehat{\text{Err}}^{\text{train}}(\varphi_{\mathcal{L}}) = \overline{E}(\varphi_{\mathcal{L}}, \mathcal{L}) = \frac{1}{N'} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{L}} L(y_i, \varphi_{\mathcal{L}}(\mathbf{x}_i)) \quad (7)$$

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- La primer (y más sencilla) forma de estimar el error de generalización es la estimación por resustitución o estimación del conjunto de entrenamiento.
- Consiste en estimar empíricamente $\text{Err}(\varphi_{\mathcal{L}})$ con el mismo conjunto de datos \mathcal{L} que se utiliza para construir $\varphi_{\mathcal{L}}$:

$$\widehat{\text{Err}}^{\text{train}}(\varphi_{\mathcal{L}}) = \overline{E}(\varphi_{\mathcal{L}}, \mathcal{L}) = \frac{1}{N'} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{L}} L(y_i, \varphi_{\mathcal{L}}(\mathbf{x}_i)) \quad (7)$$

- Esta estimación es mala, ya que ofrece un resultado optimista del error generalizado ya que usa TODOS los datos (\mathbf{x}_i, y_i) en \mathcal{L} .

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- La primer (y más sencilla) forma de estimar el error de generalización es la estimación por resustitución o estimación del conjunto de entrenamiento.
- Consiste en estimar empíricamente $\text{Err}(\varphi_{\mathcal{L}})$ con el mismo conjunto de datos \mathcal{L} que se utiliza para construir $\varphi_{\mathcal{L}}$:

$$\widehat{\text{Err}}^{\text{train}}(\varphi_{\mathcal{L}}) = \overline{E}(\varphi_{\mathcal{L}}, \mathcal{L}) = \frac{1}{N'} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{L}} L(y_i, \varphi_{\mathcal{L}}(\mathbf{x}_i)) \quad (7)$$

- Esta estimación es mala, ya que ofrece un resultado optimista del error generalizado ya que usa TODOS los datos (\mathbf{x}_i, y_i) en \mathcal{L} .
- Mundánamente, evaluar el error con todo el conjunto de datos.

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- La segunda forma de estimar el error de generalización es la estimación del conjunto de prueba.

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- La segunda forma de estimar el error de generalización es la estimación del conjunto de prueba.
- Consiste partir \mathcal{L} en dos conjuntos disjuntos $\mathcal{L}_{\text{train}}$ y $\mathcal{L}_{\text{test}}$, los conjuntos de entrenamiento y prueba, respectivamente.

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- La segunda forma de estimar el error de generalización es la estimación del conjunto de prueba.
- Consiste partir \mathcal{L} en dos conjuntos disjuntos $\mathcal{L}_{\text{train}}$ y $\mathcal{L}_{\text{test}}$, los conjuntos de entrenamiento y prueba, respectivamente.
- Por lo tanto, el error de generalización se estima usando el error promedio sobre el conjunto de prueba con el modelo creado con el conjunto de entrenamiento:

$$\widehat{\text{Err}}^{\text{test}}(\varphi_{\mathcal{L}}) = \overline{E}(\varphi_{\mathcal{L}_{\text{train}}}, \mathcal{L}_{\text{test}}) = \frac{1}{N'} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{L}_{\text{test}}} L(y_i, \varphi_{\mathcal{L}_{\text{train}}}(\mathbf{x}_i)) \quad (8)$$

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- La segunda forma de estimar el error de generalización es la estimación del conjunto de prueba.
- Consiste partir \mathcal{L} en dos conjuntos disjuntos $\mathcal{L}_{\text{train}}$ y $\mathcal{L}_{\text{test}}$, los conjuntos de entrenamiento y prueba, respectivamente.
- Por lo tanto, el error de generalización se estima usando el error promedio sobre el conjunto de prueba con el modelo creado con el conjunto de entrenamiento:

$$\widehat{\text{Err}}^{\text{test}}(\varphi_{\mathcal{L}}) = \overline{E}(\varphi_{\mathcal{L}_{\text{train}}}, \mathcal{L}_{\text{test}}) = \frac{1}{N'} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{L}_{\text{test}}} L(y_i, \varphi_{\mathcal{L}_{\text{train}}}(\mathbf{x}_i)) \quad (8)$$

- Mundánamente, realizar la partición prueba-entrenamiento, entrenar con el de entrenamiento, y determinar el error con el de prueba.

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- Debemos tener cuidado con que $\mathcal{L}_{\text{train}}$ y $\mathcal{L}_{\text{test}}$ sean independientes uno del otro.

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- Debemos tener cuidado con que $\mathcal{L}_{\text{train}}$ y $\mathcal{L}_{\text{test}}$ sean independientes uno del otro.
- Además, deben de determinarse con la misma distribución.

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- Debemos tener cuidado con que $\mathcal{L}_{\text{train}}$ y $\mathcal{L}_{\text{test}}$ sean independientes uno del otro.
- Además, deben de determinarse con la misma distribución.
- Lo cual se verifica si se eligen al azar de \mathcal{L} .

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- Debemos tener cuidado con que $\mathcal{L}_{\text{train}}$ y $\mathcal{L}_{\text{test}}$ sean independientes uno del otro.
- Además, deben de determinarse con la misma distribución.
- Lo cual se verifica si se eligen al azar de \mathcal{L} .
- Esta estrategia genera un error sin sesgo.

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- Debemos tener cuidado con que $\mathcal{L}_{\text{train}}$ y $\mathcal{L}_{\text{test}}$ sean independientes uno del otro.
- Además, deben de determinarse con la misma distribución.
- Lo cual se verifica si se eligen al azar de \mathcal{L} .
- Esta estrategia genera un error sin sesgo.
- Sin embargo, esta estrategia reduce el tamaño del conjunto para entrenar el modelo y, en consecuencia, puede que esta estrategia no estime correctamente el error de generalización si \mathcal{L} es pequeño.

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- Si \mathcal{L} es pequeño, se puede estimar mediante validación cruzada con K pliegues.

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- Si \mathcal{L} es pequeño, se puede estimar mediante validación cruzada con K pliegues.
- Divide \mathcal{L} en K subconjuntos disjuntos $\mathcal{L}_1, \dots, \mathcal{L}_K$.

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- Si \mathcal{L} es pequeño, se puede estimar mediante validación cruzada con K pliegues.
- Divide \mathcal{L} en K subconjuntos disjuntos $\mathcal{L}_1, \dots, \mathcal{L}_K$.
- Para obtener una estimación del error de generalización, promedia el error de predicción sobre los pliegues \mathcal{L}_K de los modelos $\varphi_{\mathcal{L}/\mathcal{L}_K}$ entrenado con los datos restantes:

$$\widehat{\text{Err}}^{\text{cv}}(\varphi_{\mathcal{L}}) = \frac{1}{K} \sum_{k=1}^K \overline{E}(\varphi_{\mathcal{L}/\mathcal{L}_k}, \mathcal{L}_K) \quad (9)$$

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- Si \mathcal{L} es pequeño, se puede estimar mediante validación cruzada con K pliegues.
- Divide \mathcal{L} en K subconjuntos disjuntos $\mathcal{L}_1, \dots, \mathcal{L}_K$.
- Para obtener una estimación del error de generalización, promedia el error de predicción sobre los pliegues \mathcal{L}_K de los modelos $\varphi_{\mathcal{L}/\mathcal{L}_K}$ entrenado con los datos restantes:

$$\widehat{\text{Err}}^{\text{cv}}(\varphi_{\mathcal{L}}) = \frac{1}{K} \sum_{k=1}^K \overline{E}(\varphi_{\mathcal{L}/\mathcal{L}_k}, \mathcal{L}_K) \quad (9)$$

- Ya que cada modelo $\varphi_{\mathcal{L}/\mathcal{L}_K}$ se entrena con casi todo \mathcal{L} , todos deben acercarse al modelo $\varphi_{\mathcal{L}}$ que se entrena con todo \mathcal{L} .

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- En consecuencia, las estimaciones $\overline{E}(\varphi_{\mathcal{L}/\mathcal{L}_K}, \mathcal{L}_K)$ deben ser cercanas a $\text{Err}(\varphi_{\mathcal{L}})$

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- En consecuencia, las estimaciones $\overline{E}(\varphi_{\mathcal{L}/\mathcal{L}_K}, \mathcal{L}_K)$ deben ser cercanas a $\text{Err}(\varphi_{\mathcal{L}})$
- Es más caro computacionalmente, la validación cruzada con K pliegues tiene la ventaja de que utiliza todos los puntos $(\bar{x}, y) \in \mathcal{L}$ para estimar $\text{Err}(\varphi_{\mathcal{L}})$.

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

- En consecuencia, las estimaciones $\overline{E}(\varphi_{\mathcal{L}/\mathcal{L}_K}, \mathcal{L}_K)$ deben ser cercanas a $\text{Err}(\varphi_{\mathcal{L}})$
- Es más caro computacionalmente, la validación cruzada con K pliegues tiene la ventaja de que utiliza todos los puntos $(\bar{x}, y) \in \mathcal{L}$ para estimar $\text{Err}(\varphi_{\mathcal{L}})$.
- El valor de K suele ser 10, el cual permite obtener estimaciones estables y confiables [1].

Evaluación de desempeño de modelos

Estimación de $\text{Err}(\varphi_{\mathcal{L}})$

Definición: Error de generalización esperado

Como hemos mostrado, nuestro objetivo es estimar el error de generalización $\text{Err}(\varphi_{\mathcal{L}})$ condicionado al conjunto de aprendizaje \mathcal{L} . Una cantidad relacionada es el error de generalización esperado:

$$\mathbb{E}_{\mathcal{L}}[\text{Err}(\varphi_{\mathcal{L}})] \quad (10)$$

el cual promedia todo lo que sea aleatorio, incluido el ruido que tiene el conjunto \mathcal{L} que se utiliza para entrenar $\varphi_{\mathcal{L}}$. Esta cantidad es cercana, pero diferente, a $\text{Err}(\varphi_{\mathcal{L}})$. Sin embargo, lo que las estimaciones anteriores estiman es la Ec. (10) más que la Ec. (1).

Evaluación de desempeño de modelos

Modelo de Bayes y Error Residual

- Si la distribución $P(X, Y)$ se conoce, el mejor modelo, i.e., el modelo φ_B el error de generalización de la Ec. (1) se puede derivar analíticamente e independiente del conjunto de datos \mathcal{L} .

Evaluación de desempeño de modelos

Modelo de Bayes y Error Residual

- Si la distribución $P(X, Y)$ se conoce, el mejor modelo, i.e., el modelo φ_B el error de generalización de la Ec. (1) se puede derivar analíticamente e independiente del conjunto de datos \mathcal{L} .
- Condicionando sobre X , el error de generalización se puede reescribir como

$$\mathbb{E}_{X,Y}[L(Y, \varphi_B(x))] = \mathbb{E}_X[\mathbb{E}_{Y|X}[L(Y, \varphi_B(x))]] \quad (11)$$

Evaluación de desempeño de modelos

Modelo de Bayes y Error Residual

- Si la distribución $P(X, Y)$ se conoce, el mejor modelo, i.e., el modelo φ_B el error de generalización de la Ec. (1) se puede derivar analíticamente e independiente del conjunto de datos \mathcal{L} .
- Condicionando sobre X , el error de generalización se puede reescribir como

$$\mathbb{E}_{X,Y}[L(Y, \varphi_B(x))] = \mathbb{E}_X[\mathbb{E}_{Y|X}[L(Y, \varphi_B(x))]] \quad (11)$$

- El modelo que minimiza la Ec. (11) es aquel que minimiza el valor esperado interno punto a punto:

$$\varphi_B(\mathbf{x}) = \operatorname{argmin}_{y \in \mathcal{Y}} \mathbb{E}_{Y|X=\mathbf{x}}[L(Y, y)] \quad (12)$$

Evaluación de desempeño de modelos

Modelo de Bayes y Error Residual

- El modelo φ_B se le conoce como **modelo de Bayes** y su error de generalización asociado $\text{Err}(\varphi_B)$ como **error residual**.

Evaluación de desempeño de modelos

Modelo de Bayes y Error Residual

- El modelo φ_B se le conoce como **modelo de Bayes** y su error de generalización asociado $\text{Err}(\varphi_B)$ como **error residual**.
- Representa el mínimo error que cualquier modelo de aprendizaje supervisado puede alcanzar.

Evaluación de desempeño de modelos

Modelo de Bayes y Error Residual

- El modelo φ_B se le conoce como **modelo de Bayes** y su error de generalización asociado $\text{Err}(\varphi_B)$ como **error residual**.
- Representa el mínimo error que cualquier modelo de aprendizaje supervisado puede alcanzar.
- Es el error irreducible debido puramente a desviaciones aleatorias en los datos.

Evaluación de desempeño de modelos

Modelo de Bayes y Error Residual

Definición: Modelo de Bayes

Un modelo φ_B es un modelo de Bayes si, para cualquier modelo φ entrenado con cualquier conjunto de datos \mathcal{L} , $\text{Err}(\varphi_B) \leq \text{Err}(\varphi_{\mathcal{L}})$

Evaluación de desempeño de modelos

Modelo de Bayes y Error Residual

Para clasificación, cuando \mathcal{L} es la pérdida cero-uno, el modelo de Bayes es:

$$\begin{aligned}\varphi_B(\mathbf{x}) &= \operatorname{argmin}_{y \in \mathcal{Y}} \mathbb{E}_{Y|X=\mathbf{x}}[L(Y, y)] \\ &= \operatorname{argmin}_{y \in \mathcal{Y}} \mathbb{E}_{Y|X=\mathbf{x}}[1(Y, y)] \\ &= \operatorname{argmin}_{y \in \mathcal{Y}} P(Y \neq y | X = \mathbf{x}) \\ &= \operatorname{argmax}_{y \in \mathcal{Y}} P(Y = y | X = \mathbf{x})\end{aligned}\tag{13}$$

Es decir, el mejor clasificador posible consiste en clasificar sistemáticamente la clase más probable $y \in \{c_1, \dots, c_n\}$ dado $X = \mathbf{x}$.

Evaluación de desempeño de modelos

Modelo de Bayes y Error Residual

Para regresión, con el error cuadrático, el modelo de Bayes es:

$$\begin{aligned}\varphi_B(\mathbf{x}) &= \operatorname{argmin}_{y \in \mathcal{Y}} \mathbb{E}_{Y|X=\mathbf{x}}[L(Y, y)] \\ &= \operatorname{argmin}_{y \in \mathcal{Y}} \mathbb{E}_{Y|X=\mathbf{x}}[(Y - y)^2] \\ &= \mathbb{E}_{Y|X=\mathbf{x}}[Y]\end{aligned}\tag{14}$$

En otras palabras, el mejor regresor posible consiste en sistemáticamente predecir el valor promedio de Y en $X = \mathbf{x}$.

Evaluación de desempeño de modelos

Modelo de Bayes y Error Residual

- En problemas prácticos, $P(X, Y)$ es desconocida, y el modelo de Bayes no se puede determinar analíticamente.

Evaluación de desempeño de modelos

Modelo de Bayes y Error Residual

- En problemas prácticos, $P(X, Y)$ es desconocida, y el modelo de Bayes no se puede determinar analíticamente.
- En este contexto, la eficacia del modelo $\varphi_{\mathcal{L}}$ puede ser difícil de evaluar ya que estimaciones de $\text{Err}(\varphi_{\mathcal{L}})$ pueden no ser indicativas de la bondad de $\varphi_{\mathcal{L}}$ si se desconoce el error más bajo posible $\text{Err}(\varphi_B)$.

Evaluación de desempeño de modelos

Modelo de Bayes y Error Residual

- En problemas prácticos, $P(X, Y)$ es desconocida, y el modelo de Bayes no se puede determinar analíticamente.
- En este contexto, la eficacia del modelo $\varphi_{\mathcal{L}}$ puede ser difícil de evaluar ya que estimaciones de $\text{Err}(\varphi_{\mathcal{L}})$ pueden no ser indicativas de la bondad de $\varphi_{\mathcal{L}}$ si se desconoce el error más bajo posible $\text{Err}(\varphi_B)$.
- Si se llegará a conocer φ_B , entonces podríamos comparar el error del modelo con la estimación del conjunto de prueba o validación cruzada.

Descomposición en Sesgo y Varianza

Recordemos que definimos el error de predicción esperado en la Ec. (1) como:

$$\text{Err}(\varphi_{\mathcal{L}}) = \mathbb{E}_{X,Y}[L(Y, \varphi_{\mathcal{L}}(X))]$$

Además, el error de predicción esperado de $\varphi_{\mathcal{L}}$ en $X = \mathbf{x}$ se puede expresar como:

$$\text{Err}(\varphi_{\mathcal{L}}(\mathbf{x})) = \mathbb{E}_{Y|X=\mathbf{x}}[L(Y, \varphi_{\mathcal{L}}(\mathbf{x}))] \quad (15)$$

En regresión, para el error cuadrático, esta última forma del error de predicción esperado se descompone aditivamente en términos de sesgo y varianza que, en conjunto, constituyen un marco muy útil para el diagnóstico del error de predicción de un modelo.

Descomposición en Sesgo y Varianza

En regresión, suponiendo que \mathcal{L} es el error cuadrático, el error de predicción esperado de un modelo $\varphi_{\mathcal{L}}$ en un punto dado $X = \mathbf{x}$ se puede reescribir con respecto al modelo de Bayes φ_B :

$$\begin{aligned}\text{Err}(\varphi_{\mathcal{L}}(\mathbf{x})) &= \mathbb{E}_{Y|X=\mathbf{x}}[(Y - \varphi_{\mathcal{L}}(\mathbf{x}))^2] \\ &= \mathbb{E}_{Y|X=\mathbf{x}}[(Y - \varphi_B(\mathbf{x}) + \varphi_B(\mathbf{x}) - \varphi_{\mathcal{L}}(\mathbf{x}))^2]\end{aligned}$$

Descomposición en Sesgo y Varianza

En regresión, suponiendo que \mathcal{L} es el error cuadrático, el error de predicción esperado de un modelo $\varphi_{\mathcal{L}}$ en un punto dado $X = \mathbf{x}$ se puede reescribir con respecto al modelo de Bayes φ_B :

$$\begin{aligned}\text{Err}(\varphi_{\mathcal{L}}(\mathbf{x})) &= \mathbb{E}_{Y|X=\mathbf{x}}[(Y - \varphi_{\mathcal{L}}(\mathbf{x}))^2] \\ &= \mathbb{E}_{Y|X=\mathbf{x}}[(Y - \varphi_B(\mathbf{x}) + \varphi_B(\mathbf{x}) - \varphi_{\mathcal{L}}(\mathbf{x}))^2]\end{aligned}$$

Ejercicio

Terminar de desarrollar la expresión anterior.

Descomposición en Sesgo y Varianza

En regresión, suponiendo que \mathcal{L} es el error cuadrático, el error de predicción esperado de un modelo $\varphi_{\mathcal{L}}$ en un punto dado $X = \mathbf{x}$ se puede reescribir con respecto al modelo de Bayes φ_B :

$$\begin{aligned}\text{Err}(\varphi_{\mathcal{L}}(\mathbf{x})) &= \mathbb{E}_{Y|X=\mathbf{x}}[(Y - \varphi_{\mathcal{L}}(\mathbf{x}))^2] \\&= \mathbb{E}_{Y|X=\mathbf{x}}[(Y - \varphi_B(\mathbf{x}) + \varphi_B(\mathbf{x}) - \varphi_{\mathcal{L}}(\mathbf{x}))^2] \\&= \mathbb{E}_{Y|X=\mathbf{x}}[(Y - \varphi_B(\mathbf{x}))^2] + \mathbb{E}_{Y|X=\mathbf{x}}[(\varphi_B(\mathbf{x}) - \varphi_{\mathcal{L}}(\mathbf{x}))^2] \\&\quad \hookrightarrow + \mathbb{E}_{Y|X=\mathbf{x}}[2(Y - \varphi_B(\mathbf{x}))(\varphi_B(\mathbf{x}) - \varphi_{\mathcal{L}}(\mathbf{x}))] \\&= \mathbb{E}_{Y|X=\mathbf{x}}[(Y - \varphi_B(\mathbf{x}))^2] + \mathbb{E}_{Y|X=\mathbf{x}}[(\varphi_B(\mathbf{x}) - \varphi_{\mathcal{L}}(\mathbf{x}))^2] \\&= \text{Err}(\varphi_B(\mathbf{x})) + (\varphi_B(\mathbf{x}) - \varphi_{\mathcal{L}}(\mathbf{x}))^2\end{aligned}\tag{16}$$

ya que $\mathbb{E}_{Y|X=\mathbf{x}}[Y - \varphi_B(\mathbf{x})] = \mathbb{E}_{Y|X=\mathbf{x}}[Y] - \varphi_B(\mathbf{x}) = 0$ por la Eq. (14).

Descomposición en Sesgo y Varianza

$$\text{Err}(\varphi_B(\mathbf{x})) + (\varphi_B(\mathbf{x}) - \varphi_{\mathcal{L}}(\mathbf{x}))^2$$

El primer término en la Ec. (16) corresponder al error residual irreducible en $X = \mathbf{x}$, mientras que el segundo representa la discrepancia entre $\varphi_{\mathcal{L}}$ y el modelo de Bayes.

Descomposición en Sesgo y Varianza

- Si asumimos además que el conjunto de datos \mathcal{L} es en sí mismo una variable aleatoria (muestreado de la población ω) y que el algoritmo de aprendizaje es determinista ...
- ... entonces la discrepancia esperada sobre \mathcal{L} con el modelo de Bayes se puede volver a expresar en términos de la predicción promedio $\mathbb{E}_{\mathcal{L}}[\varphi_{\mathcal{L}}(\mathbf{x})]$ sobre los modelos aprendidos de todos los posibles conjuntos de aprendizaje de tamaño N .

Descomposición en Sesgo y Varianza

$$\begin{aligned}\mathbb{E}_{\mathcal{L}}[(\varphi_B(\mathbf{x}) - \varphi_{\mathcal{L}}(\mathbf{x}))^2] &= \mathbb{E}_{\mathcal{L}}[(\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}[\varphi_{\mathcal{L}}(\mathbf{x})] + \mathbb{E}_{\mathcal{L}}[\varphi_{\mathcal{L}}(\mathbf{x})] - \varphi_{\mathcal{L}}(\mathbf{x}))^2] \\&= \mathbb{E}_{\mathcal{L}}[(\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}[\varphi_{\mathcal{L}}(\mathbf{x})])^2] + \mathbb{E}_{\mathcal{L}}[(\mathbb{E}_{\mathcal{L}}[\varphi_{\mathcal{L}}(\mathbf{x})] - \varphi_{\mathcal{L}}(\mathbf{x}))^2] \\&\hookrightarrow +\mathbb{E}_{\mathcal{L}}[2(\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}[\varphi_{\mathcal{L}}(\mathbf{x})])(\mathbb{E}_{\mathcal{L}}[\varphi_{\mathcal{L}}(\mathbf{x})] - \varphi_{\mathcal{L}}(\mathbf{x}))] \\&= \mathbb{E}_{\mathcal{L}}[(\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}[\varphi_{\mathcal{L}}(\mathbf{x})])^2] + \mathbb{E}_{\mathcal{L}}[(\mathbb{E}_{\mathcal{L}}[\varphi_{\mathcal{L}}(\mathbf{x})] - \varphi_{\mathcal{L}}(\mathbf{x}))^2] \\&= (\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}[\varphi_{\mathcal{L}}(\mathbf{x})])^2 + \mathbb{E}_{\mathcal{L}}[(\mathbb{E}_{\mathcal{L}}[\varphi_{\mathcal{L}}(\mathbf{x})] - \varphi_{\mathcal{L}}(\mathbf{x}))^2]\end{aligned}\tag{17}$$

ya que $\mathbb{E}_{\mathcal{L}}[\mathbb{E}_{\mathcal{L}}[\varphi_{\mathcal{L}}(\mathbf{x})] - \varphi_{\mathcal{L}}(\mathbf{x})] = \mathbb{E}_{\mathcal{L}}[\varphi_{\mathcal{L}}(\mathbf{x})] - \mathbb{E}_{\mathcal{L}}[\varphi_{\mathcal{L}}(\mathbf{x})] = 0$

Descomposición en Sesgo y Varianza

Teoream

Para el error cuadrático medio, la descomposición en sesgo y varianza del valor esperado del error de generalización $\mathbb{E}_{\mathcal{L}}[\text{Err}(\varphi_{\mathcal{L}}(\mathbf{x}))]$ en $X = \mathbf{x}$ es

$$\mathbb{E}_{\mathcal{L}}[\text{Err}(\varphi_{\mathcal{L}}(\mathbf{x}))] = \text{noise}(\mathbf{x}) + \text{bias}^2(\mathbf{x}) + \text{var}(\mathbf{x}) \quad (18)$$

donde

$$\text{noise}(\mathbf{x}) = \text{Err}(\varphi_b(\mathbf{x})),$$

$$\text{bias}^2(\mathbf{x}) = (\varphi_B(\mathbf{x}) - \mathbb{E}_{\mathcal{L}}[\varphi_{\mathcal{L}}(\mathbf{x})])^2,$$

$$\text{var}(\mathbf{x}) = \mathbb{E}_{\mathcal{L}}[(\mathbb{E}_{\mathcal{L}}[\varphi_{\mathcal{L}}(\mathbf{x})] - \varphi_{\mathcal{L}}(\mathbf{x}))^2]$$

Descomposición en Sesgo y Varianza

- El primer término, $\text{noise}(\mathbf{x})$, es el **error residual**. Es independiente del modelo de aprendizaje y el conjunto de datos y provee una cota inferior en el error de generalización.
- El segundo término, $\text{bias}^2(\mathbf{x})$ mide la **discrepancia** entre la predicción promedio y la predicción del modelo de Bayes.
- El tercer término, $\text{var}(\mathbf{x})$, mide la **variabilidad** de las predicciones en $X = \mathbf{x}$ sobre los modelos aprendidos de todos los conjuntos de aprendizaje posibles.

Descomposición en Sesgo y Varianza

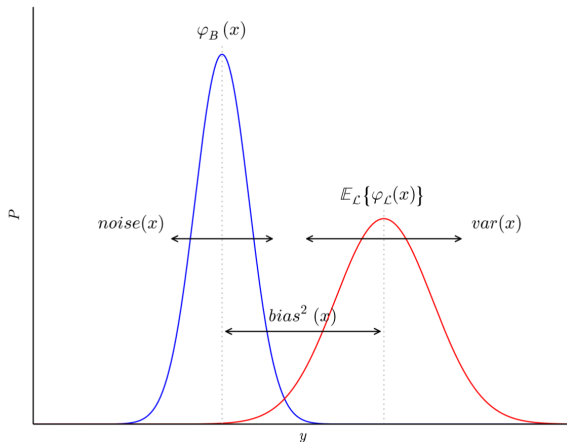


Figure 4.1: Residual error, bias and variance at $X = x$. (Figure inspired from [Geurts, 2002].)

Figura: En la imagen, tanto el ruido como la varianza miden la dispersión de ambas densidades. El sesgo mide la distancia entre las media.

Descomposición en Sesgo y Varianza

- ¿Qué tiene que ver esto con la navidad?

Descomposición en Sesgo y Varianza

- ¿Qué tiene que ver esto con la navidad?
- El teorema anterior revela el rol de la varianza en error de generalización esperado.

Descomposición en Sesgo y Varianza

- ¿Qué tiene que ver esto con la navidad?
- El teorema anterior revela el rol de la varianza en error de generalización esperado.
- Por lo tanto, una forma de reducir el error sería reducir el de alguno de sus componentes, que resultan ser el sesgo y la varianza.

Descomposición en Sesgo y Varianza

- ¿Qué tiene que ver esto con la navidad?
- El teorema anterior revela el rol de la varianza en error de generalización esperado.
- Por lo tanto, una forma de reducir el error sería reducir el de alguno de sus componentes, que resultan ser el sesgo y la varianza.
- Vamos a enfocarnos en reducir en la varianza (asumiendo que el sesgo no cambie demasiado).

Descomposición en Sesgo y Varianza

- ¿Qué tiene que ver esto con la navidad?
- El teorema anterior revela el rol de la varianza en error de generalización esperado.
- Por lo tanto, una forma de reducir el error sería reducir el de alguno de sus componentes, que resultan ser el sesgo y la varianza.
- Vamos a enfocarnos en reducir en la varianza (asumiendo que el sesgo no cambie demasiado).
- Esto lo permite hacer los ensambles. La aleatorización que introduce provoca perturbaciones en el proceso de aprendizaje ya que jugamos con el conjunto de datos \mathcal{L} .

Tarea

Leer la tesis de Louppe [2], específicamente el capítulo 4.1, 4.2 (incluidas todas las sub secciones).

Nota: Parte de eso vendrá en el examen.

- [1] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, page 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.
- [2] Gilles Louppe. Understanding random forests: From theory to practice, 2014.