Evaluación de Modelos (I)

Carlos Hurtado L.

Departamento de Ciencias de la Computación, U de Chile

Evaluación de Desempeño: Decisiones a tomar

- Datos de prueba
- Medidas de efectividad
- Problema típico: predecir desempeño cuando los datos son limitados.
- Estrategia para reducir sesgo
- Considerar vs. no considerar costo de errores
- Evaluación de un modelo vs. comparación de varios modelos

Otras dimensiones de desempeño

- En esta sección solo estudiaremos desempeño como capacidad predictiva.
- Otras dimensiones de un modelo:
  - Tiempo que toma en clasificar
  - Costo de construcción
  - Tamaño
  - Interpretabilidad (por personas)

Digresión sobre estadística

Referencia:

Introductory Statistics. Thomas Wonnacott, Ronald Wonnacott, John Wiley. 1990.

Tarea: lectura de capítulos 1-10.

Contenido (Cap. 1-10)

- Nociones básicas de probabilidades
- Variables Aleatorias
- Distribuciones de probabilidad.
- Muestreo
- Estimadores
- Intervalos de Confianza
- Verificación de Hipótesis
- Análisis de la Varianza (ANOVA)

Muestreo

- Disciplina que estudia cómo obtener conclusiones de una población a partir de una muestra de datos de ella.

Población

- Población: colección completa de objetos que estamos estudiando.
  - Por ejemplo, si intentamos predecir una elección, la población es el universo de votantes.
- Típicamente, estudiamos como se comporta una variable X en la población.
  - En el ejemplo anterior, la variable sería el candidato por el que vota cada persona.

Recordemos: en clasificación

- Datos de entrenamiento (DE)
  - Datos que se usan para entrenar el modelo
- **Muestra**: Datos de prueba (DP)
  - Datos que se usan para probar el modelo
- **Población**: Datos objetivo (DO)
  - Datos sobre los cuales se ejecuta el modelo

Variable a estudiar

- Variable aleatoria X con distribución de probabilidades p(x).
- Media: $\mu := E(X) = \sum x p(x)$
- Varianza:

$$\sigma^2 := \mathrm{var}(X) := E((X-\mu)^2)$$

$$\sigma^2 = \sum (x-\mu)^2 p(x)$$

$$\sigma^2 = \sum x^2 p(x) - \mu^2$$

Muestra

- Subconjunto de la población.
- Muestra aleatoria (población discreta):
  - Asignamos un número entero a cada objeto.
  - Seleccionamos números aleatoriamiente.
  - Cada vez que elegimos un número podemos reemplazarlo o no.

Muestra Aleatoria Muy Simple (MAMS)

- Obtenemos una observación Xi por cada objeto seleccionado.
- Las observaciones Xi’s son independientes y cada uno distribuye p(x) (como la población)
- Es decir, cada observación tiene media μ y varianza σ

MAMS y Reemplazo

- Si la población es pequeña y obtenemos una muestra generada sin reemplazo no es una MAMS.
- Si la poblacióne es muy grande, sin reemplazo sí podemos obtener una MAMS (aproximadamente).

Media de muestra

- La media de muestra $\bar{X}$ es una variable aleatoria cuyo dominio son las medias de todas las MAMS de tamaño n.
- Media de $\bar{X}$

$$E(\bar{X}) = \frac{1}{n}\sum E(X_i) = \mu$$

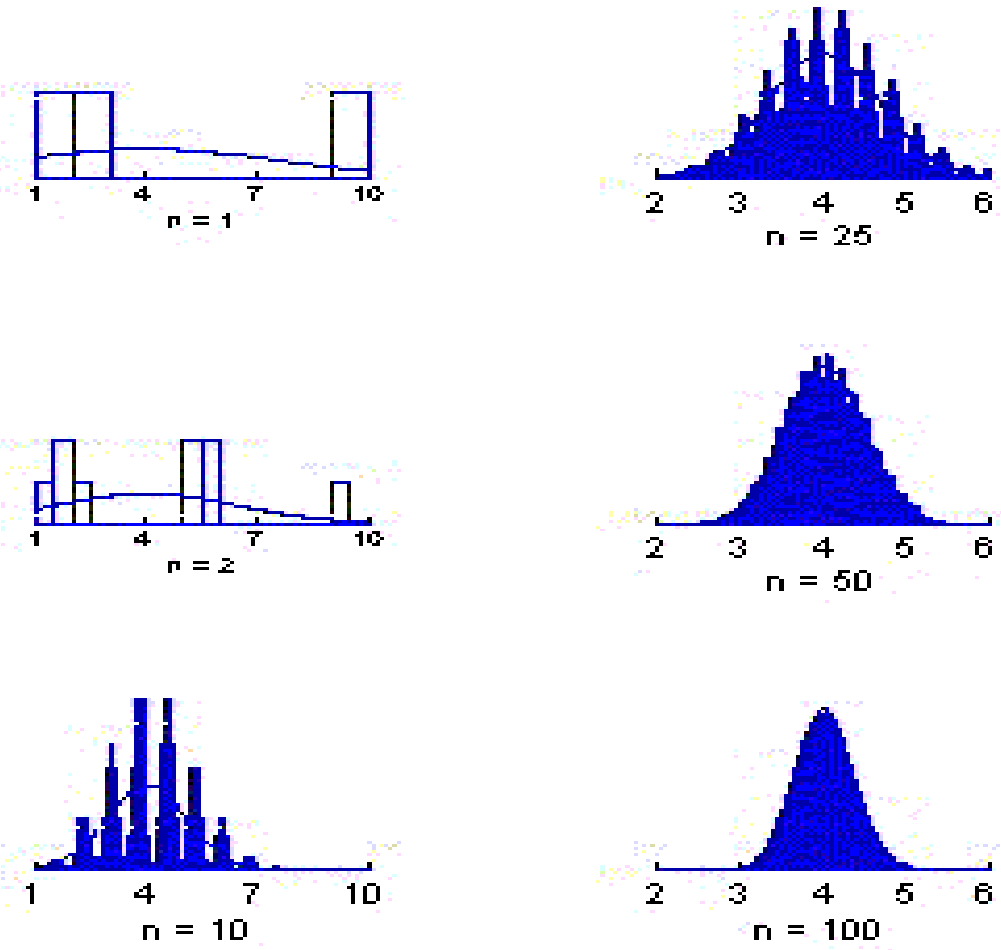
- Varianza de $\bar{X}$

$$\text{var}(\bar{X}) = E((\bar{X}-\mu)^2) = \text{var}(\frac{1}{n}\sum X_i) = \frac{1}{n^2}\sum \text{var}(X_i) = \frac{\sigma^2}{n}$$

- Error estándar de

$$SE = \frac{\sigma}{\sqrt{n}}$$

Distribución de la Media de Muestra



Teorema del Límite Central

- Si n es grande entonces la media de muestra distribuye aproximadamente normal.
- Esto se aplica independientemente de si la población distribuye normal.
- Es decir a medida que n crece, tenemos que

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- Adicionalmente, a medida que n crece, el error estándar disminuye (tiende a cero)

Proporciones

- Es el caso que tenemos cuando X es binaria (1,0).
- La media de X la denotamos π y representa el % de la población para la cual X=1.
- La varianza de X es $\pi(1-\pi)$
- En este caso la “media de muestra” se denomina “proporción de la muestra”.

Proporción de la muestra

- La proporción de la muestra $\bar{X}$ es una variable aleatoria cuyo dominio son las proporciones de todas las muestras de tamaño n.
- Media de $\bar{X}$ $E(\bar{X})=\pi$
- Varianza de $\bar{X}$ $\text{var}(\bar{X})=\frac{\pi(1-\pi)}{n}$
- Error estándar $SE=\sqrt{\frac{\pi(1-\pi)}{n}}$

Ejemplo

- ¿De nuestros 15 primeros nietos cuál es la chance que tengamos más de 10 hombre?
- En este caso sabemos que $\pi = 0.5$
- Luego $\bar{X} \sim \mathrm{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$
- La respuesta es

$$P\left(\bar{X} > \frac{10}{15}\right) = 0.99$$

Muestreo sin reemplazo

- Para una muestra con reemplazo tenemos

$$SE = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

- Es decir, sin reemplazo, podemos reducir el error dependiendo de los tamaños de la muestra y la población.

Intervalos de Confianza

- Debido a que $\bar{X} \sim \mathrm{N}(\mu, \mathrm{SE}^2)$
- Podemos definir un intervalo con 95% confianza μ para

$$\mu = \bar{X} \pm z_{.025} SE$$

$$\mu = \bar{X} \pm z_{.025} \frac{\sigma}{\sqrt{n}}$$

Varianza desconocida

- Si no conocemos la varianza y n es pequeño (n < 100) en la práctica, tenemos

$$\mu = \bar{X} \pm t_{.025} AproxSE$$

$$\mu = \bar{X} \pm t_{.025} \frac{s}{\sqrt{n}}$$

$$s = \frac{1}{n-1} \sum (X_i - \bar{X})^2$$

Intervalos de Confianza para proporciones

- Debido a que $\bar{X} \sim \mathrm{N}(\pi, \frac{\pi(1-\pi)}{n})$
- Podemos definir un intervalo con 95% confianza π para

$$\pi = \bar{X} \pm z_{.025} SE$$

$$\pi = \bar{X} \pm z_{.025} \sqrt{\frac{\pi(1-\pi)}{n}}$$

Método de Montecarlo

- La idea del método de Montecarlo es seleccionar muchas muestras, calcular para c/u $\bar{X}$ e intentar reproducir $p(\bar{x})$
- Se puede construir un intervalo para μ en un rango 95% central de la distribución.
- Es útil cuando no podemos aproximar a distribución normal y podemos obtener suficientes muestras (ejemplo, en simulación).

Bootstrap

- A partir de la muestra reproducimos la población. Por ejemplo replicamos 1000 veces la muestra.
- Luego usamos el método de Montecarlo para construir un intervalo para μ
- Es útil cuando no podemos aproximar a distribución normal y tenemos una única muestra.

Bootstrap: Ejemplo

- Considere una muestra aleatoria de n=10 observaciones:

  M= 16,12,14,6,43,7,0,54,25,13

- Construimos (virtualmente) la población “Bootstrap” replicando cada dato un millon de veces.

Bootstrap: Ejemplo (cont.)

- Luego, obtenemos muestras de 10 datos.
- Cada muestra se obtiene eligiendo aleatoriamente de M con reemplazo (esto equivale a seleccionar muestras de la población Bootsrap).
- Después de calcular el para 1000 muestras obtenemos un rango (95% central):

$\bar{X}$

$$9 < \mu < 26$$

Boostrap (cont.)

- El intervalo Boostrap:

$$9 < \mu < 26$$

- Es similar a un intervalo

$$\mu = \bar{X} \pm t_{.025} \frac{s}{\sqrt{n}}$$

$$7 < \mu < 31$$

Estrategías para Evaluar Desempeño

- Resustitución y Holdout
- Validación Cruzada
- Funciones de pérdida
- Contabilización del Costo
- Métodos Gráficos
- Comparación de Modelos

Terminología

- Datos de entrenamiento (DE)
  - Datos que se usan para entrenar el modelo
- Muestra: Datos de prueba (DP)
  - Datos que se usan para probar el modelo
- Población: Datos objetivo (DO)
  - Datos sobre los cuales se ejecuta el modelo

Estimación Simple de Desempeño

- Medida de efectividad:
  - *Error de predicción*: fracción de datos mal clasificados en datos objetivo.
  - Equivalente a *1 - tasa de éxito*
- Tasa de error se estima en base a:
  - **"Resustitución"**: estimamos el error en datos de entrenamiento
  - **"Holdout"**: estimamos el error en datos de prueba disjuntos de datos de entrenamiento

Varianza en la Estimación del Error Ejemplo: weather.nominal

- Supongamos que inducimos la regla

```
If outlook=rainy then play=yes
Otherwise play=no
```

- Se evalúa en el siguiente conjunto:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	High	FALSE	No

- Tenemos tasa de error igual a cero

Varianza en la Estimación del Error Ejemplo: weather.nominal (cont.)

- Ahora inducimos la regla

```
If outlook=rainy then play=yes
Otherwise play=no
```

- Esta vez se evalúa en el siguiente conjunto:

Outlook	Temp.	Humidity	Windy	Play
Overcast	Hot	High	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Cool	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

- Ahora tenemos tasa de error igual a 100%

Intervalo de confianza para el Error

- Supongamos que observamos *s* errores en un conjunto de *n* datos
- Estimamos el error como la proporción de la muestra

$$\bar{X} = \frac{s}{n}$$

- Se tiene

$$\bar{X} \sim \mathrm{N}\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

Intervalo de confianza (IC) para error

$$\pi = \bar{X} \pm z_{.025}\sqrt{\frac{\pi(1-\pi)}{n}}$$

$$\pi = \frac{\bar{X} + \frac{z^2}{2n} \pm z\sqrt{\frac{\bar{x}}{n} - \frac{\bar{x}}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}$$

Resustitución

- Estimación de error sobre datos de entrenamiento
- En este caso el error estimado se denomina “error de resustitución”
- Problemas:
  - El modelo tiende a mostrar un buen desempeño en datos de entrenamiento
  - Mala predicción de error sobre datos objetivo

Holdout

- Típicamente se tiene un conjunto de datos disponibles
- Dividir datos disponibles en dos conjuntos disjuntos: DE y DP
- Se requiere:
  - Los dos conjuntos (construcción y prueba) sean **representativos** de los datos objetivos
- Dilema:
  - Para producir un buen clasificador necesitamos usar la mayor cantidad de datos para entrenamiento
  - Para obtener una buena estimación del error objetivo necesitamos usar la mayor cantidad de datos para prueba

Datos representativos para Holdout

- En general, no se puede saber si los datos de entrenamiento o prueba son representativos.
- Sí podemos saber si están todas las clases “representadas”
  - Si una clase no está representada en los datos de entrenamiento, el modelo no tendrá un buen desempeño para la clase.
  - Si la clase no está representada en los datos de entrenamiento, no mediremos bien el error asociado a la clase.

Técnicas para evitar sesgo en Holdout

- Holdout estratificado:
  - Clases ocurren con la misma frecuencia en partición entrenamiento/prueba.
  - Salvaguarda básica para sesgo.
- Holdout repetitivo:
  - Repetir la prueba varias veces pero cambiando la partición entrenamiento/prueba.
  - Error estimado: promedio de errores de cada iteración

Validación cruzada (“cross validation”)

- Forma de Hold-out repetitivo
- Validación cruzada de “**n**-fold”
  - Datos se dividen en un número **n** fijo de subconjuntos
  - Dado un subconjunto s, se usa s como prueba y los datos restantes como entrenamiento.
  - Esto se repite para cada subconjunto

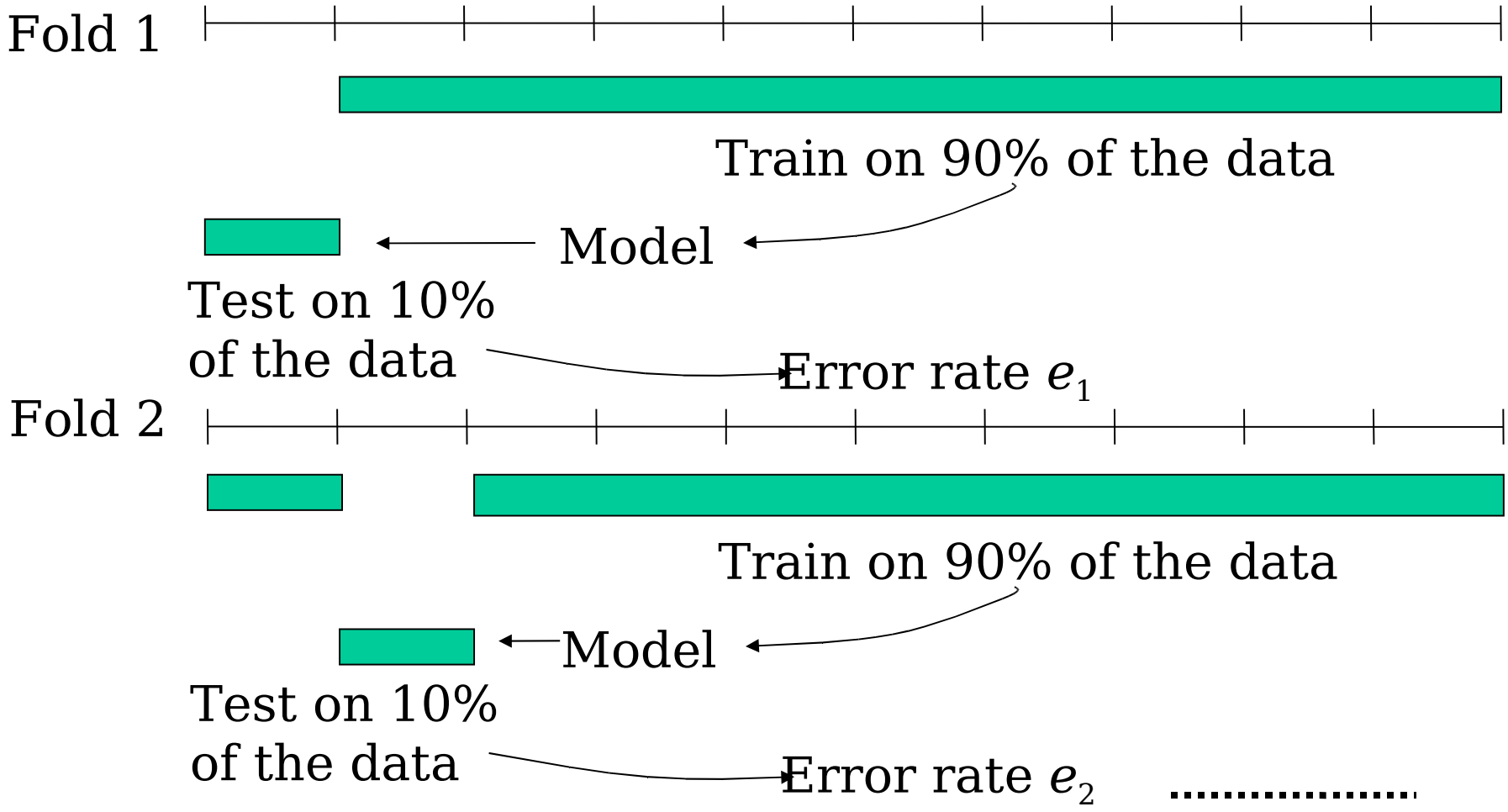
Validación cruzada (“cross validation”)

- Error estimado: promedio de los errores en cada iteración
- Se puede usar estratificación
- Desventaja: costo computacional. Se debe inducir el modelo n veces.
  - No es factible para conjuntos de datos grandes.

Validación cruzada (“cross validation”) (cont.)

- Uso típico: 10 veces validación cruzada de 10-fold
- “leave-one-out”: caso particular, donde n es número de datos
  - útil cuando se tienen pocos datos.

Validación Cruzada



Estimación del Error en Validación Cruzada

- Estimación del error

$$\bar{Y} = \frac{1}{k} \sum_{i=1}^{k} \bar{X}_k$$

- La variable $\bar{Y}$ tiene media π

- Desviación estandar de $\bar{Y}$

$$s_{\bar{Y}^2} = \frac{1}{(k-1)} \sum_{i=1}^{k} (\bar{X}_k - \bar{Y})^2$$

Estimación del Error en Validación Cruzada (cont.)

- Tenemos $$\frac{\bar{Y}-\pi}{\sqrt{\frac{s_{\bar{Y}^2}}{k}}} \sim t_{k-1}$$

- Obtenemos el intervalo (95% de confianza):

$$\pi = \bar{Y} \pm t_{.025} \frac{s_{\bar{Y}}}{\sqrt{k}}$$

Ejercicio

- Corra J48 con weather y test de prueba igual a datos de entrenamiento
- Corra J48 con weather y percentage split de 66%
- Corra J48 con weather y 10-fold cross validation
- Compare tasa de error.