Clasificación II

Rubén Francisco Manrique
rf.manrique@uniandes.edu.co

Clasificación de texto

- Asignar categorías, tópicos, géneros.
- Detección de spam.
- Identificación de autoría.
- Identificación de edad/genero.
- Identificación de lenguaje
- Análisis de sentimientos.

Overfitting (Sobreajuste)

- Un modelo que coincide perfectamente con los datos de entrenamiento tiene un problema.
- También se sobreajustará a los datos, modelando el ruido.
  - Una palabra aleatoria que predice perfectamente y (sólo ocurre en una clase) obtendrá una ponderación muy alta.
  - No poder generalizar a un conjunto de pruebas sin esta palabra.

Un buen modelo debería poder **generalizar**

Overfitting

Training data

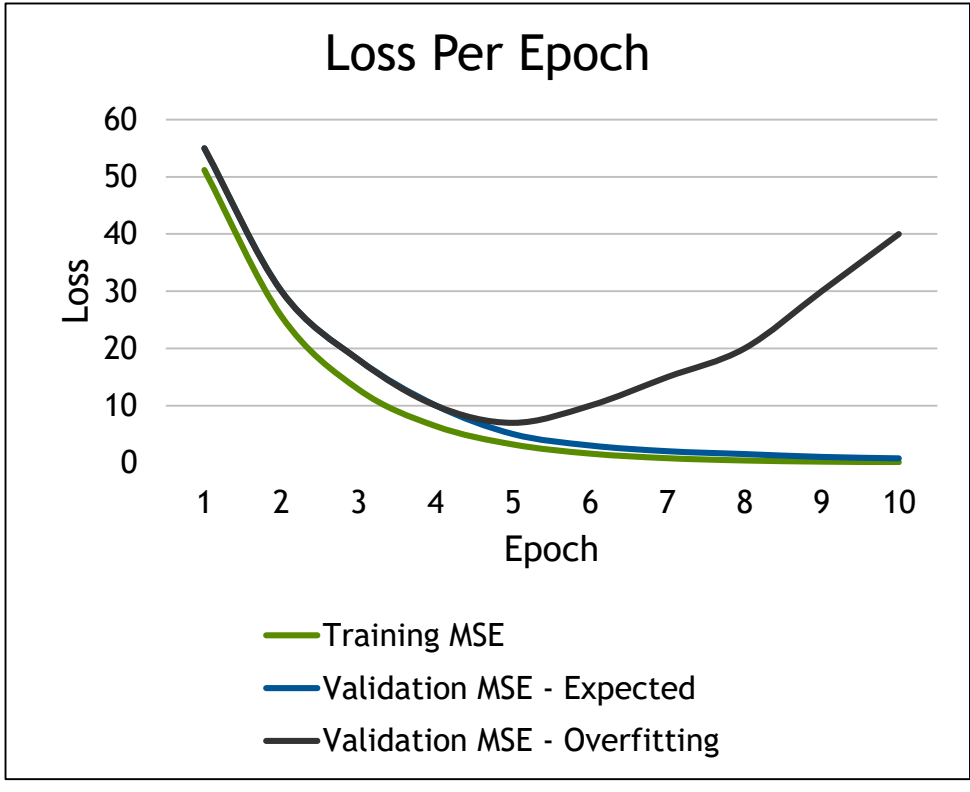
- Core dataset for the model to learn on

Validation data

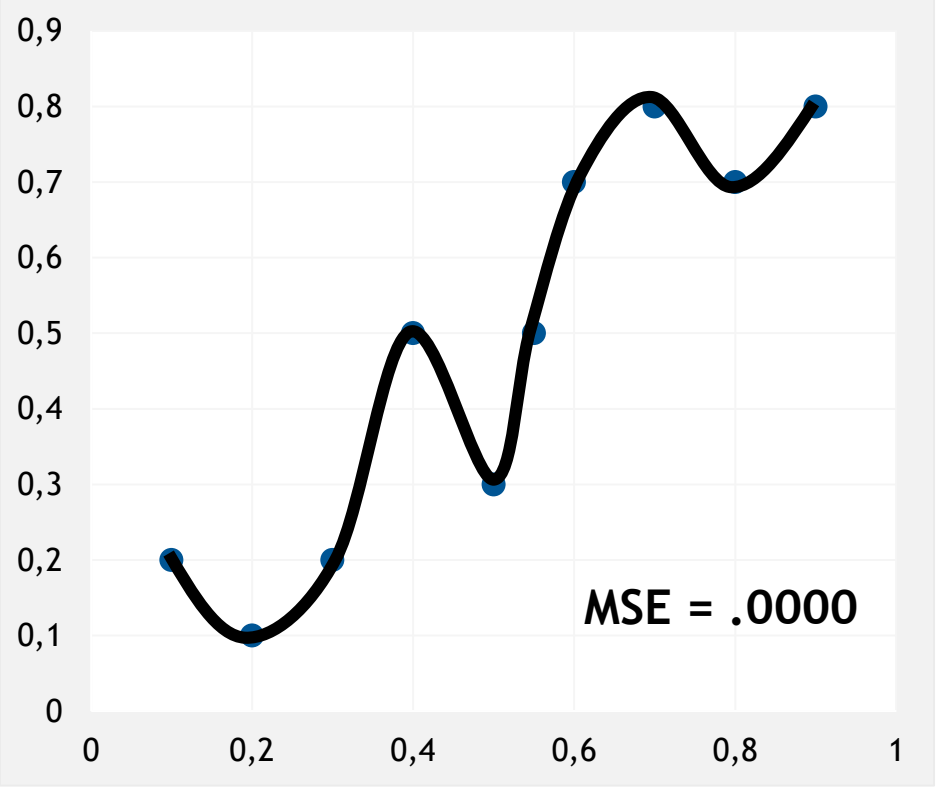
- New data for model to see if it truly understands (can generalize)

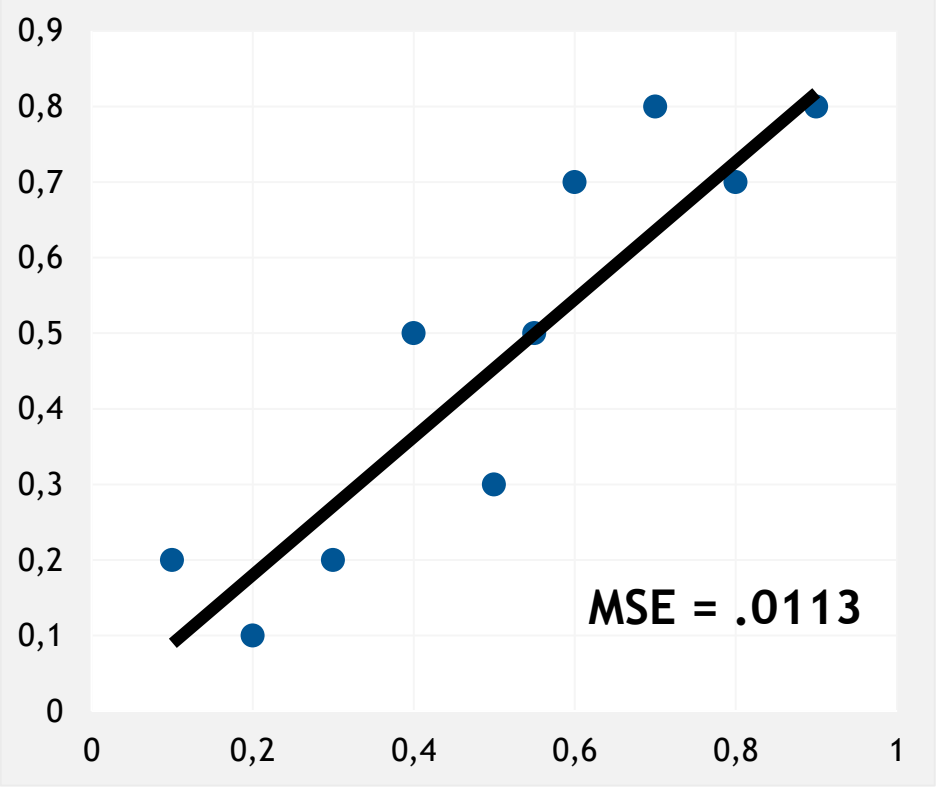
Overfitting

- When model performs well on the training data, but not the validation data (evidence of memorization)
- Ideally the accuracy and loss should be similar between both datasets

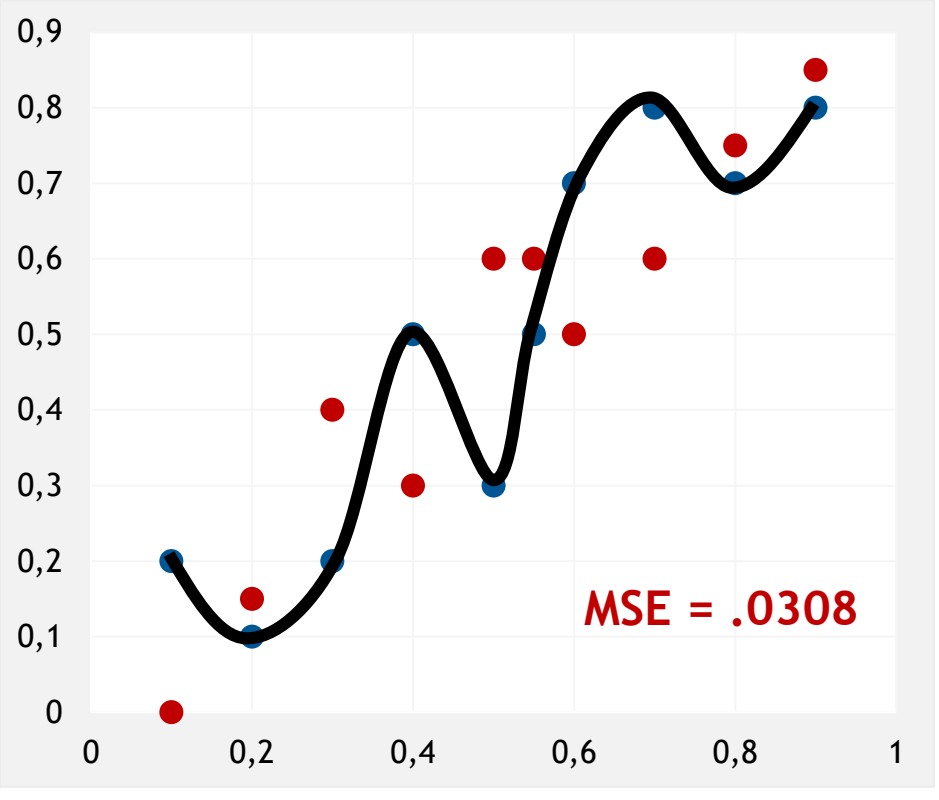


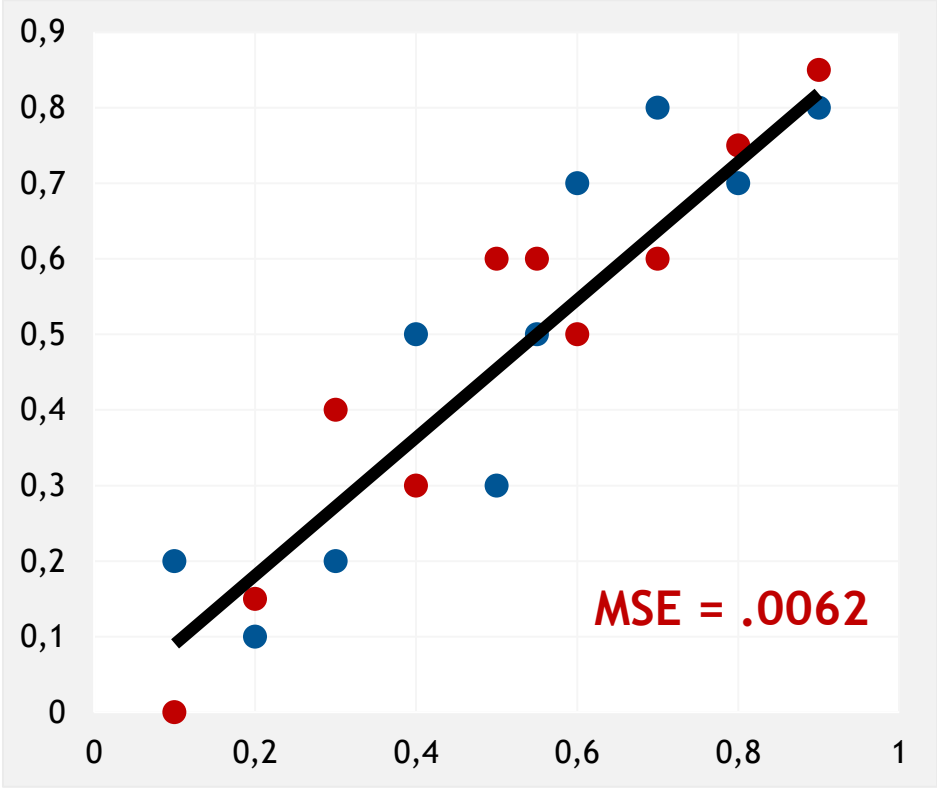
Overfitting (Training)





Overfitting (Testing)





Regularización

- Adicionar una función de regularización que penalice pesos grandes.
  - Ajustar bien los datos con ponderaciones grandes (pesos) no es tan bueno como ajustar los datos un poco peor, con ponderaciones (pesos) pequeñas.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{m} \log P(y^{(i)}|x^{(i)}) - \alpha R(\theta)$$

Regularización L2 (ridge regression)

- Idea: La suma de cuadrados de sus pesos

$$R(\theta) = ||\theta||_2^2 = \sum_{j=1}^{n} \theta_j^2$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \left[\sum_{i=1}^{m} \log P(y^{(i)} | x^{(i)}) \right] - \alpha \sum_{j=1}^{n} \theta_j^2$$

Regularización L1 (lasso regression)

- Idea: La suma de cuadrados de sus pesos

$$R(\theta) = ||\theta||_1 = \sum_{i=1}^{n} |\theta_i|$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \left[\sum_{i=1}^{m} \log P(y^{(i)} | x^{(i)}) \right] - \alpha \sum_{j=1}^{n} \theta_j^2$$

Como lo hacemos multinomial

- Frecuentemente se necesitan más de dos clases:
  - Positivo/negativo/neutral
  - POS (sustantivos, verbos, adjetivos, adverbios)
- Si >2 multinomial logistic regression
  - Multinomial logic
  - Maximun entropy
- Tenemos el siguiente reto:

$$P(\text{positive}|\text{doc}) + P(\text{negative}|\text{doc}) + P(\text{neutral}|\text{doc}) = 1$$

Multinomial análisis

- La entrada sigue siendo el producto escalar entre el vector de peso w y el vector de entrada x
- Pero ahora necesitaremos vectores de peso separados para cada una de las K clases.

- Supongamos tres clases (+,-,0). Cada feature va a tener asociado 3 pesos.

Feature	Definition	$w_{5,+}$	$w_{5,-}$	$w_{5,0}$
$f_5(x)$	$\begin{cases} 1 & \text{if } \text{“!”} \in \text{doc} \\ 0 & \text{otherwise} \end{cases}$	3.5	3.1	−5.3

La función softmax

- En últimas vamos a tener K valores de probabilidad, uno por cada clase:
  - $P(+|x), P(-|x), P(0|x)$
- En regresión logística antes de estimar la probabilidad tenemos los valores z
  - $z_+ = w_+ x + b_+$
  - $z_- = w_- x + b_-$
  - $z_0 = w_0 x + b_0$
- Ya la función sigmoidal no nos sirve!
- Solución función softmax!!

La función softmax

- Torna un vector $\boldsymbol{z} = [z_1, z_2, \ldots, z_k]$ de k clases arbitrarias en probabilidades.

$$\text{softmax}(z_i) = \frac{\exp(z_i)}{\sum_{j=1}^{k} \exp(z_j)} \quad 1 \leq i \leq k$$

$$z = [0.6, 1.1, -1.5, 1.2, 3.2, -1.1]$$

$$\text{softmax}(z) = \left[\frac{\exp(z_1)}{\sum_{i=1}^{k} \exp(z_i)}, \frac{\exp(z_2)}{\sum_{i=1}^{k} \exp(z_i)}, \ldots, \frac{\exp(z_k)}{\sum_{i=1}^{k} \exp(z_i)} \right]$$

$$[0.055, 0.090, 0.0067, 0.10, 0.74, 0.010]$$

Otros aspectos prácticos

- El descenso del gradiente estocástico (SGD) elige un solo ejemplo aleatorio a la vez.
  - Eso puede resultar en movimientos entrecortados.
  - No se aprovecha la paralelización del cálculo de vectores en GPUs.
- **Mini-batch training**: *m* examples (4, 8, 16, 32, 64, 128, 512, or 1024)
- Se promedian los gradients de cada batch y con el promedio se actualizan los pesos.

$$\frac{\partial Cost(\hat{y}, y)}{\partial w_j} = \frac{1}{m} \sum_{i=1}^{m} \left[\sigma(\mathbf{w} \cdot \mathbf{x}^{(i)} + b) - y^{(i)} \right] \mathbf{x}_j^{(i)}$$

Evaluación

The 2-by-2 contingency table (Confusion Matriz)

		Truth/Real Value	
		spam	not spam
Predicted	spam	tp	fp
	not spam	fn	tn

Precision and recall

- **Precision**: % of selected items that are correct
  Recall: % of correct items that are selected

$$P = tp/(tp + fp)$$
$$R = tp/(tp + fn)$$

		Truth/Real Value	
		spam	not spam
Predicted	spam	tp	fp
	not spam	fn	tn

A combined measure: F

- A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

$$F = \frac{1}{a\frac{1}{P} + (1-a)\frac{1}{R}} = \frac{(b^2+1)PR}{b^2P+R}$$

- The harmonic mean is a very conservative average; see *IIR* § 8.3
- People usually use balanced F1 measure
  - i.e., with $\beta = 1$ (that is, $\alpha = ½$): $F1 = 2PR/(P+R)$

¿Multinomial?
Classic Reuters-21578 Data Set

- Most (over)used data set, 21,578 docs (each 90 types, 200 tokens)
- 9603 training, 3299 test articles (ModApte/Lewis split)
- 118 categories
  - An article can be in more than one category
  - Learn 118 binary category distinctions
- Average document (with at least one category) has 1.24 classes
- Only about 10 out of 118 categories are large

Common categories
(#train, #test)

- Earn (2877, 1087)
- Acquisitions (1650, 179)
- Money-fx (538, 179)
- Grain (433, 149)
- Crude (389, 189)
- Trade (369,119)
- Interest (347, 131)
- Ship (197, 89)
- Wheat (212, 71)
- Corn (182, 56)

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" NEWID="798">

<DATE> 2-MAR-1987 16:51:43.42</DATE>

<TOPICS><D>livestock</D><D>hog</D></TOPICS>

<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>

<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.

Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter

&#3;</BODY></TEXT></REUTERS>

A new confusión matrix

Docs in test set	Assigned UK	Assigned poultry	Assigned wheat	Assigned coffee	Assigned interest	Assigned trade
True UK	95	1	13	0	1	0
True poultry	0	1	0	0	0	0
True wheat	10	90	0	1	0	0
True coffee	0	0	0	34	3	7
True interest	-	1	2	13	26	5
True trade	0	0	2	14	5	10

Per class evaluation measures

Recall:

Fraction of docs in class *i* classified correctly:

$$\frac{c_{ii}}{\sum_j c_{ij}}$$

Precision:

Fraction of docs assigned class *i* that are actually about class *i*:

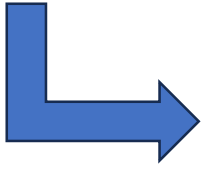
$$\frac{c_{ii}}{\sum_j c_{ji}}$$

Accuracy: (1 - error rate)

Fraction of docs classified correctly:

$$\frac{\sum_i c_{ii}}{\sum_j \sum_i c_{ij}}$$

Docs in test set	Assigned UK	Assigned poultry	Assigned wheat	Assigned coffee	Assigned interest	Assigned trade
True UK	95	1	13	0	1	0
True poultry	0	1	0	0	0	0
True wheat	10	90	0	1	0	0
True coffee	0	0	0	34	3	7
True interest	-	1	2	13	26	5
True trade	0	0	2	14	5	10



UK

	Truth: UK	Truth: no-UK
Classifier: UK	95	10
Classifier: no-UK	15	214

Micro- vs. Macro-Averaging

- If we have more than one class, how do we combine multiple performance measures into one quantity?

- **Macroaveraging**: Compute performance for each class, then average.

- **Microaveraging**: Collect decisions for all classes, compute contingency table, evaluate.

Micro- vs. Macro-Averaging: Example

Class 1

	Truth: yes	Truth: no
Classifier: yes	10	10
Classifier: no	10	970

Class 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

Micro Ave. Table

	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

- Macroaveraged precision: (0.5 + 0.9)/2 = 0.7
- Microaveraged precision: 100/120 = .83

Gracias por la atención

¿Tiene alguna pregunta?