clusterAI 2020
ciencia de datos en ingeniería industrial
UTN BA
curso I5521

clase_04: ML Strategy

Presentación: Matías Callara

Mi nivel de energía está en

0

1

2

3

agenda clase04: ML Strategy

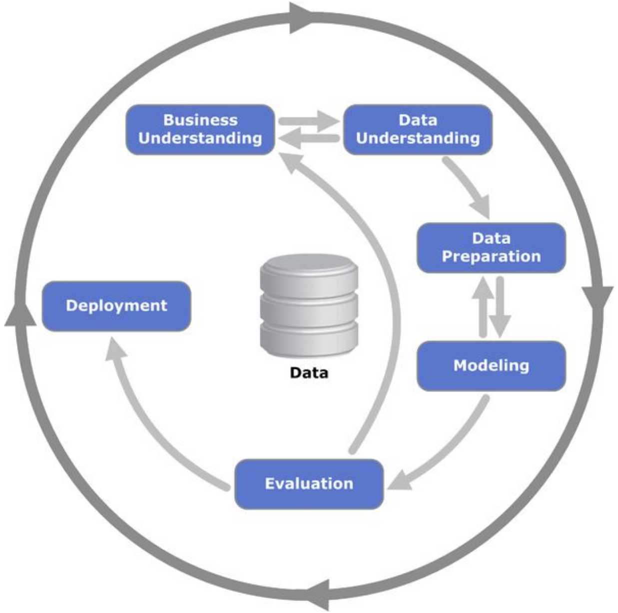
- This class is a summary of Andrew Ng's great course on ML Strategy. He is a much better communicator than me so I suggest you to go and join his course and enjoy a great teacher. 🙌

agenda clase04: ML Strategy

- Data Science Project Phases
- Orthogonalization
- Train / Dev / Test
- Training: Bias vs Variance
- Evaluation metrics

Solving problems with ML

Data Science Project Phases



Focus on one problem at a time (Orthogonalization)

Reaching the target

Use sliders a & b to reach target T.

- First, try to reach T with point NO and then with point O. Which one was easier?

Data Preparation

Train/dev/test distributions and size

Guideline

Same distribution

Choose a dev set and test set to reflect data you expect to get in the future and consider important to do well on.

Size of test set

→ Set your test set to be big enough to give high confidence in the overall performance of your system.

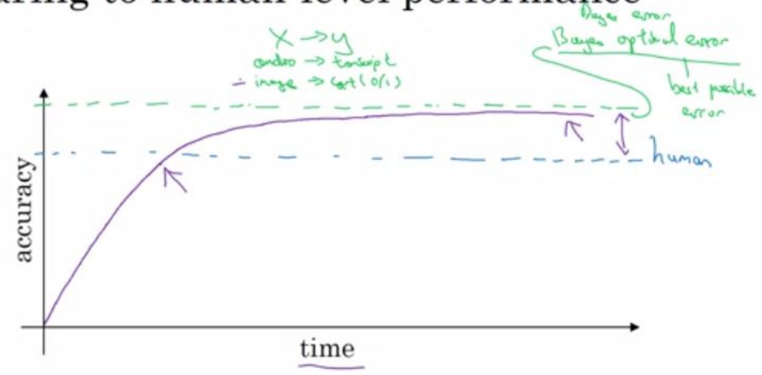
Andrew Ng

Modeling

The best possible result

Is it possible to have **perfect** performance at this task?
How can I **estimate** the best possible performance (Bayes' Error)?

Comparing to human-level performance



Andrew Ng

Why compare to human-level performance

Humans are quite good at a lot of tasks. So long as ML is worse than humans, you can:

- Get labeled data from humans. (x, y)
- Gain insight from manual error analysis: Why did a person get this right?
- Better analysis of bias/variance.

Andrew Ng

Bias vs Variance

Error Analysis. Using the training and dev error to identify on what to focus.

Training Error	8 %
Dev Error	10 %

From: Andrew Ng

Bias vs Variance

Error Analysis. Using the training and dev error to identify on what to focus.

Human Error	1 %
Training Error	8 %
Dev Error	10 %

Human error is used as a proxy for Bayes error.

From: Andrew Ng

Bias vs Variance

Error Analysis. Using the training and dev error to identify on what to focus.

Human Error	1 %
Training Error	8 %
Dev Error	10 %

(Avoidable) Bias

Human error is used as a proxy for Bayes error.

From: Andrew Ng

Bias vs Variance

Error Analysis. Using the training and dev error to identify on what to focus.

Human Error	1 %
Training Error	8 %
Dev Error	10 %

(Avoidable) Bias

Focus on **Bias**

Human error is used as a proxy for Bayes error.

From: Andrew Ng

Bias vs Variance

Error Analysis. Using the training and dev error to identify on what to focus.

Human Error	1 %	7.5 %
Training Error	8 %	8 %
Dev Error	10 %	10%
	Focus on **Bias**	

(Avoidable) Bias

Human error is used as a proxy for Bayes error.

From: Andrew Ng

Bias vs Variance

Error Analysis. Using the training and dev error to identify on what to focus.

Human Error	1 %	7.5 %
Training Error	8 %	8 %
Dev Error	10 %	10%
	Focus on **Bias**	Focus on **Variance**

(Avoidable) Bias

Variance

Human error is used as a proxy for Bayes error.

From: Andrew Ng

What is Human-level performance?

Human-level error as a proxy for Bayes error

Medical image classification example:

Suppose:

(a) Typical human .... 3 % error

(b) Typical doctor .... 1 % error

(c) Experienced doctor .... 0.7 % error

(d) Team of experienced doctors .. 0.5 % error

Bayes error ≤ 0.5%

What is "human-level" error?

Andrew Ng

Surpassing human-level Performance

Team of humans		0.5%	0.5%
One human	0.1	~~1%~~	1%
Training error		0.6%	0.3%
Dev error	0.2	0.8%	0.4%

What is avoidable bias?

Andrew Ng

Working as a Data Scientist

Get ready to compete!

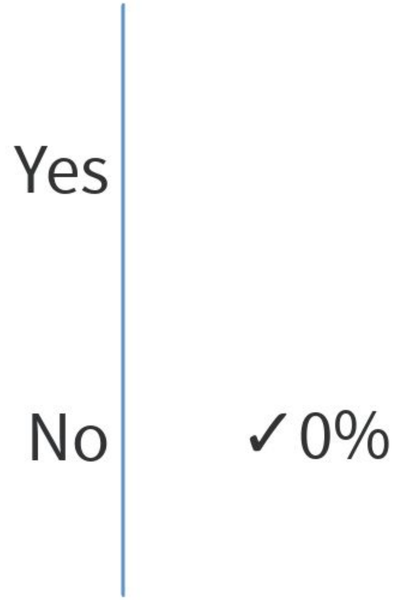
(With a lot of effort) I manage to get very small error on the training set. Am I done?

(With a lot of effort) I manage to get very small error on the training set. Am I done?

Yes

No

(With a lot of effort) I manage to get very small error on the training set. Am I done?



Leaderboard

Getting better training error

Reduce bias by

- Using a more flexible model (a bigger model)
- Train longer / use a better optimization algorithm
- Change the NN architecture / use better hyperparameter search

(With a lot of effort) I manage to get very small error on the validation set. Am I done?

When poll is active, respond at **PollEv.com/mcallara346**

(With a lot of effort) I manage to get very small error on the validation set. Am I done?

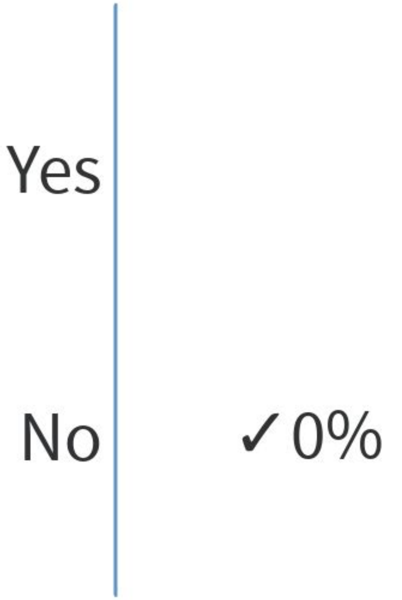
Yes

No

Tc 0

Poll locked. Responses not accepted.

(With a lot of effort) I manage to get very small error on the validation set. Am I done?



Leaderboard

Getting better dev (validation) error

Reduce variance by:

- Get more training data
- Apply regularization
- Change the NN architecture / better hyperparameter search

(With a lot of effort) I manage to get very small error on the test set. Am I done?

When poll is active, respond at **PollEv.com/mcallara346**

(With a lot of effort) I manage to get very small error on the test set. Am I done?

Yes

No

Tc 0

Poll locked. Responses not accepted.

(With a lot of effort) I manage to get very small error on the test set. Am I done?

Yes

No ✓0%

Leaderboard

Getting better test error

- Get a bigger dev dataset
- Check distributions of the dataset split

(With a lot of effort) I manage to get very small error in the real world. Am I done?

(With a lot of effort) I manage to get very small error in the real world. Am I done?

Yes

No

Tc 0

Poll locked. Responses not accepted.

(With a lot of effort) I manage to get very small error in the real world. Am I done?

Yes

No ✓0%

Leaderboard

Getting better Real Life error

- Change dev dataset
- Change cost function

Summary

The two fundamental assumptions of supervised learning

1. You can fit the training set pretty well.

   ~ Avoidable bias

2. The training set performance generalizes pretty well to the dev/test set.

   ~ Variance

Andrew Ng

Reducing (avoidable) bias and variance

Human-level

Avoidable bias →
- Train bigger model
- Train longer/better optimization algorithms
  - momentum, RMSprop, Adam
- NN architecture/hyperparameters search (RNN, CNN)

Training error

Variance →
- More data
- Regularization
  - L_2, dropout, data augmentation
- NN architecture/hyperparameters search

Dev error

Andrew Ng

Evaluation

Evaluation: Evaluation Metric vs Loss

Loss function: (Binary) Cross Entropy

$$CE = \begin{cases} & \text{if } y = 1 \\ & \end{cases}$$

Engineering Time!

Let’s create a loss function

Evaluation: Evaluation Metric vs Loss

Loss function: (Binary) Cross Entropy

$$CE = \begin{cases} -log(\hat{y}) \text{ if } y = 1 \\ \end{cases}$$

Evaluation: Evaluation Metric vs Loss

Loss function: (Binary) Cross Entropy

$$CE = \begin{cases} -log(\hat{y}) & \text{if } y = 1 \\ & \text{if } y = 0 \end{cases}$$

Evaluation: Evaluation Metric vs Loss

Loss function: (Binary) Cross Entropy

$$CE = \begin{cases} -log(\hat{y}) & \text{if } y = 1 \\ -log(1-\hat{y}) & \text{if } y = 0 \end{cases}$$

Evaluation: Evaluation Metric vs Loss

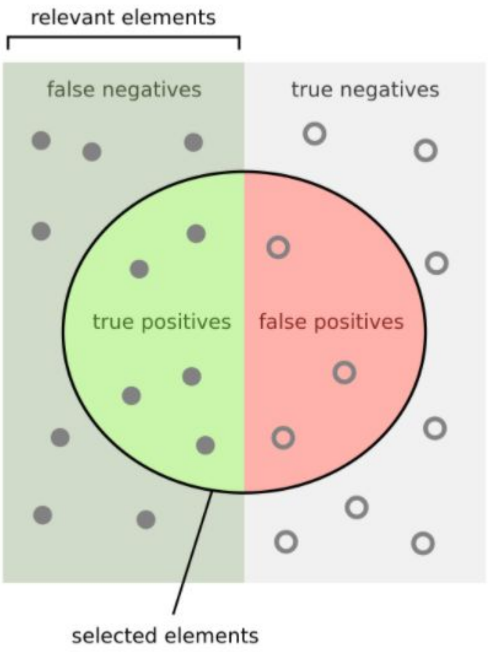
Loss function: (Binary) Cross Entropy

$$CE = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

Evaluation: Evaluation Metric vs Loss

Loss function: (Binary) Cross Entropy

$$CE = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

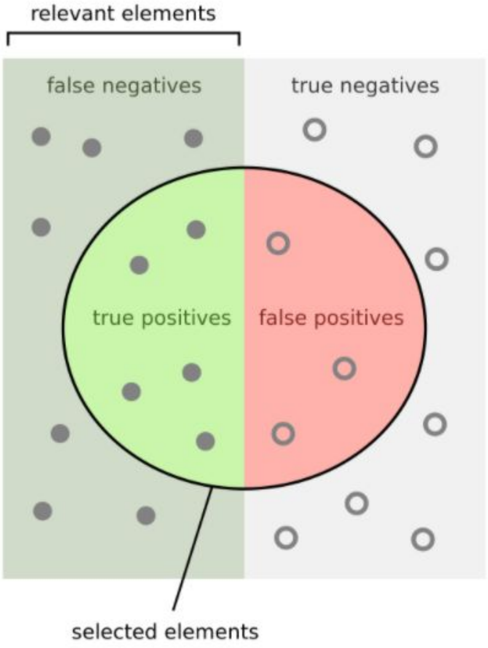


https://commons.wikimedia.org/wiki/File:Precisionrecall.svg

Evaluation: Evaluation Metric vs Loss

Loss function: (Binary) Cross Entropy

$$CE = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$



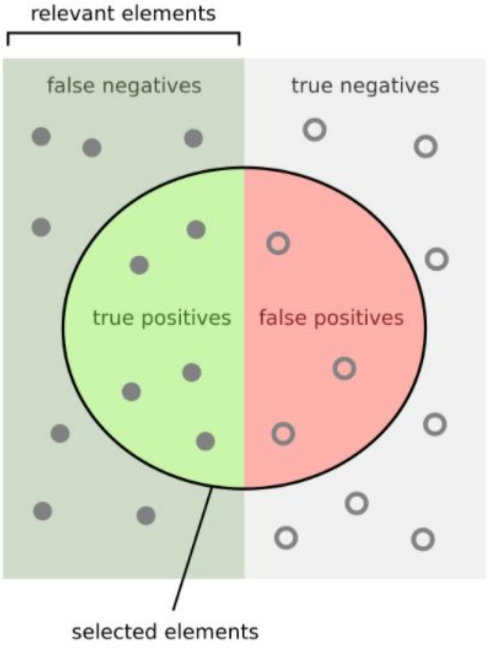
	p' (Predicted)	n' (Predicted)
p (Actual)		
n (Actual)		

https://commons.wikimedia.org/wiki/File:Precisionrecall.svg

Evaluation: Evaluation Metric vs Loss

Loss function: (Binary) Cross Entropy

$$CE = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$



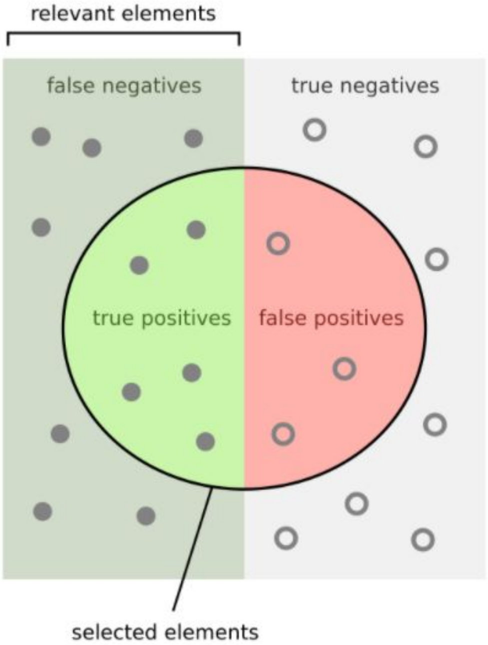
	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	
n (Actual)		

https://commons.wikimedia.org/wiki/File:Precisionrecall.svg

Evaluation: Evaluation Metric vs Loss

Loss function: (Binary) Cross Entropy

$$CE = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$



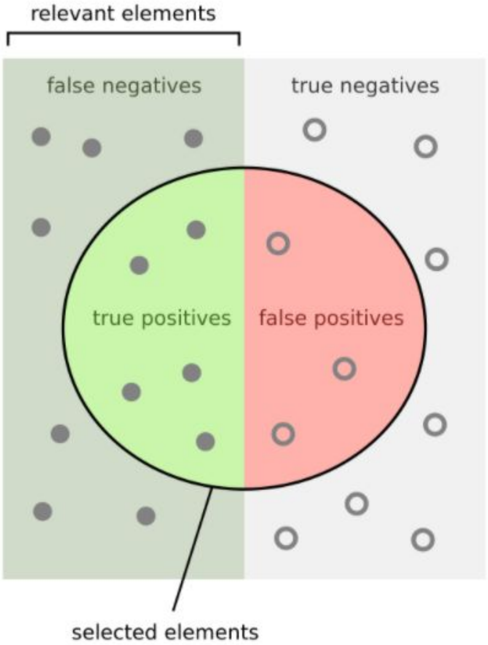
https://commons.wikimedia.org/wiki/File:Precisionrecall.svg

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)		

Evaluation: Evaluation Metric vs Loss

Loss function: (Binary) Cross Entropy

$$CE = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$



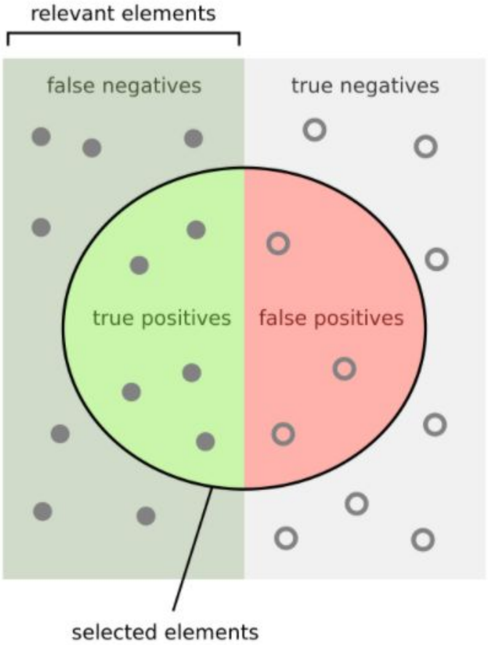
	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	

https://commons.wikimedia.org/wiki/File:Precisionrecall.svg

Evaluation: Evaluation Metric vs Loss

Loss function: (Binary) Cross Entropy

$$CE = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$



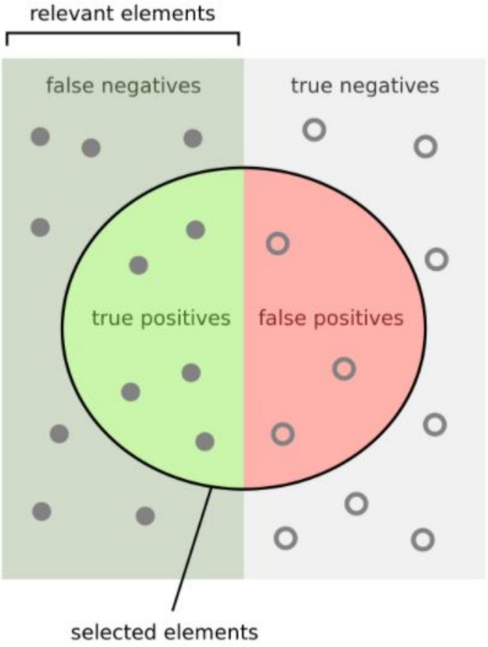
	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

https://commons.wikimedia.org/wiki/File:Precisionrecall.svg

Evaluation: Evaluation Metric vs Loss

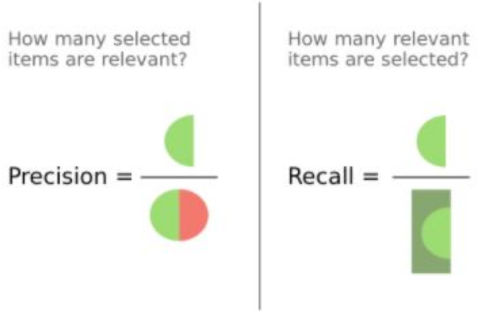
Loss function: (Binary) Cross Entropy

$$CE = -y \log(\hat{y}) - (1-y)\log(1-\hat{y})$$



https://commons.wikimedia.org/wiki/File:Precisionrecall.svg

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative



Evaluation Metrics

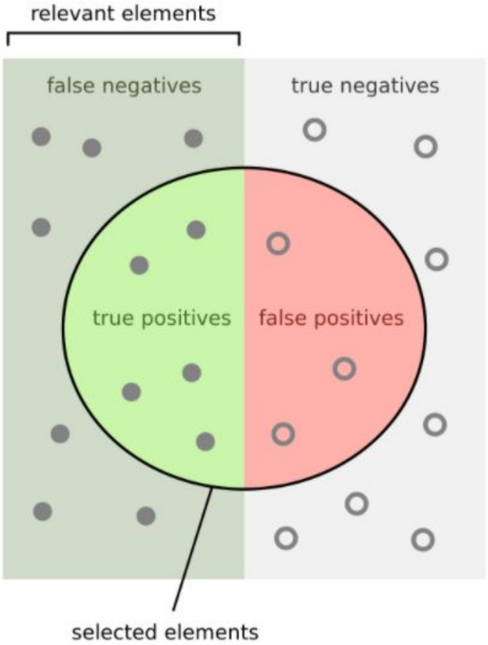
$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

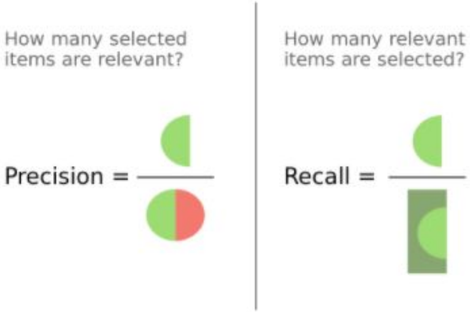
Evaluation: Evaluation Metric vs Loss

Loss function: (Binary) Cross Entropy

$$CE = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$



	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative



Evaluation Metrics

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

$$\textbf{F1-score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

https://commons.wikimedia.org/wiki/File:Precisionrecall.svg

Single Evaluation Metric

- Having a **single** real number **evaluation metric** will help you compare the performance of different models.
- You can decide how to **combine** different metrics.
- You can **change** your metric. If the models ranking generated by the metric doesn't match your personal/business ranking, it may mean that your metric is not reflecting your real objective.
- **Keep it orthogonal**. First, fix the evaluation metric then fix the system to get better at the new metric.

Satisficing and Optimizing Metrics

In the problem understanding phase we need to identify which are **objectives** we need to **optimize for** and which are **conditions** that we need **to satisfy**.

Classifier	Accuracy	Running time
A	90%	80ms
B	92%	95ms
C	95%	1,500ms

Andrew Ng

accuracy (ACC)

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Now is your turn!

https://playground.tensorflow.org/

Some thoughts...

Let's get (a bit) philosophical

In Mathematics, we try to come up with **abstractions** (structures, patterns) that will help us **deal** with **complex concepts**, **apply** them to **other situations** (context) or **compare** them with **other perspective** (abstractions) of the same phenomena.

Compare:

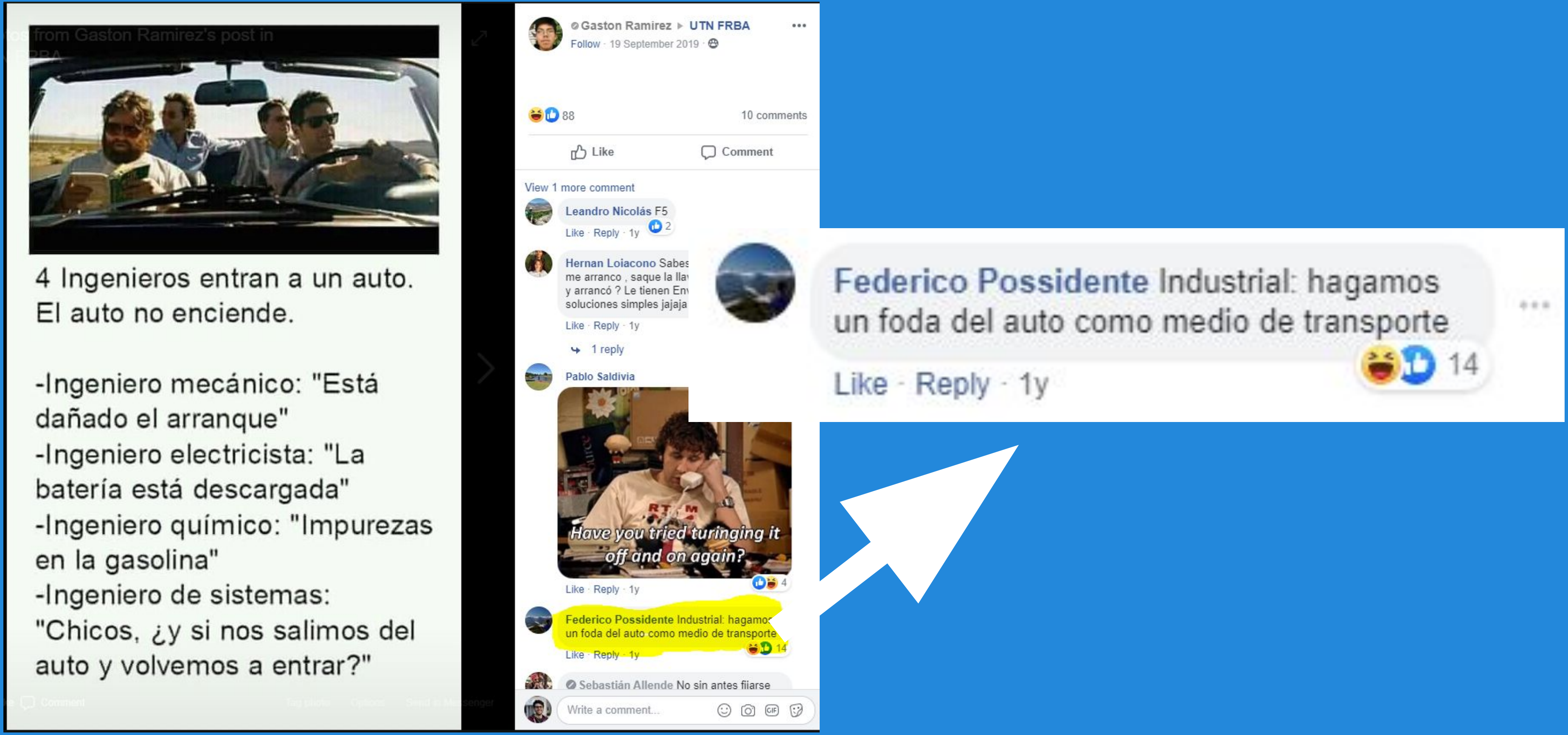
$$Ax = b$$

$$vs$$

$$\begin{cases} A_{11}x_1 + A_{12}x_2 = b1 \\ A_{21}x_1 + A_{22}x_2 = b2 \end{cases}$$

How does this relate to ML?

https://www.facebook.com/photo.php?fbid=10211102465031995&set=gm.1260906740625159&type=3&theater&ifg=1



https://www.facebook.com/photo.php?fbid=2478961752187299&set=gm.2385663964816092&type=3&theater&ifg=1

https://twitter.com/ingenieriared/status/1169178790863486981/photo/1

My experience...

- I (really) started learning mathematics after getting my engineering degree.
- At a company, I wanted to be able of solving (some) of their problems.
- Learning how to construct data-driven solutions changed the relationship with my work (and eventually my life).

For a next session...

- Relación entre funciones y vectores.
- Funciones como vectores infinitos.
- Matrices como funciones lineales.
- Operadores como “funciones” para funciones.
- Las múltiples caras de la regresión lineal (Algebraica (escalar y vectorial), Ajuste de Puntos, Geométrica, Probabilística, Logística, Red Neuronal, Ecuación Normal, Método del gradiente)