Maior Dúvida da Aula

Regularization

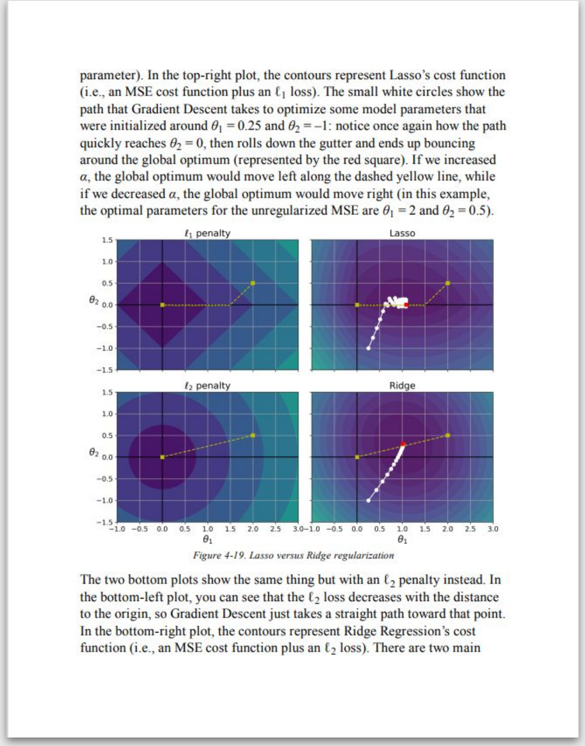
MC886/MO444, September 15, 2022

l1 ou l2?

1. Quando utilizar l1 ou l2?
    - l1 regularization penalizes the sum of absolute values of the weights,
      l2 regularization penalizes the sum of squares of the weights.
    - l1 regularization solution is sparse.
      l2 regularization solution is non-sparse.
    - l1 regularization has built-in feature selection,
      l2 regularization doesn't perform feature selection, since weights are only reduced to values near 0 instead of 0.
    - l1 regularization is robust to outliers,
      l2 regularization is not.

https://neptune.ai/blog/fighting-overfitting-with-l1-or-l2-regularization

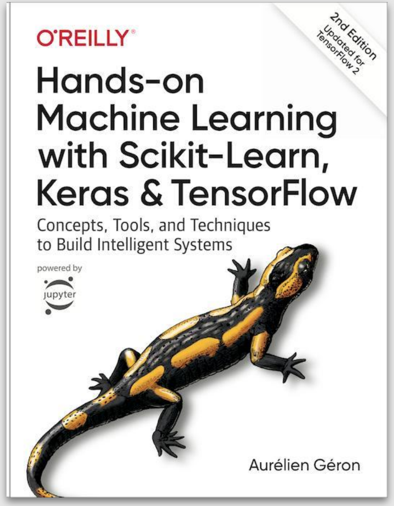
l1 ou l2?



parameter). In the top-right plot, the contours represent Lasso's cost function (i.e., an MSE cost function plus an ℓ_1 loss). The small white circles show the path that Gradient Descent takes to optimize some model parameters that were initialized around $\theta_1 = 0.25$ and $\theta_2 = -1$: notice once again how the path quickly reaches $\theta_2 = 0$, then rolls down the gutter and ends up bouncing around the global optimum (represented by the red square). If we increased α, the global optimum would move left along the dashed yellow line, while if we decreased α, the global optimum would move right (in this example, the optimal parameters for the unregularized MSE are $\theta_1 = 2$ and $\theta_2 = 0.5$).

Figure 4-19. Lasso versus Ridge regularization

The two bottom plots show the same thing but with an ℓ_2 penalty instead. In the bottom-left plot, you can see that the ℓ_2 loss decreases with the distance to the origin, so Gradient Descent just takes a straight path toward that point. In the bottom-right plot, the contours represent Ridge Regression's cost function (i.e., an MSE cost function plus an ℓ_2 loss). There are two main



O'REILLY®

2nd Edition
Updated for
TensorFlow 2

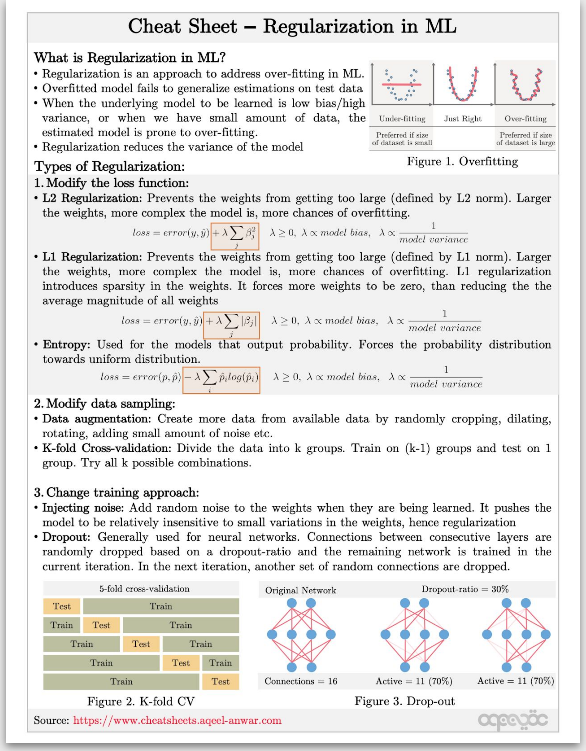
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow

Concepts, Tools, and Techniques to Build Intelligent Systems

powered by jupyter

Aurélien Géron

Chap 4, p. 201



Cheat Sheet – Regularization in ML

What is Regularization in ML?

- Regularization is an approach to address over-fitting in ML.
- Overfitted model fails to generalize estimations on test data
- When the underlying model to be learned is low bias/high variance, or when we have small amount of data, the estimated model is prone to over-fitting.
- Regularization reduces the variance of the model

Under-fitting
Just Right
Over-fitting
Preferred if size of dataset is small
Preferred if size of dataset is large

Figure 1. Overfitting

Types of Regularization:

1. Modify the loss function:

- **L2 Regularization:** Prevents the weights from getting too large (defined by L2 norm). Larger the weights, more complex the model is, more chances of overfitting.

$$loss = error(y,\hat{y}) \boxed{+\lambda \sum_j \beta_j^2} \quad \lambda \geq 0,\ \lambda \propto model\ bias,\ \ \lambda \propto \frac{1}{model\ variance}$$

- **L1 Regularization:** Prevents the weights from getting too large (defined by L1 norm). Larger the weights, more complex the model is, more chances of overfitting. L1 regularization introduces sparsity in the weights. It forces more weights to be zero, than reducing the the average magnitude of all weights

$$loss = error(y,\hat{y}) \boxed{+\lambda \sum_j |\beta_j|} \quad \lambda \geq 0,\ \lambda \propto model\ bias,\ \ \lambda \propto \frac{1}{model\ variance}$$

- **Entropy:** Used for the models that output probability. Forces the probability distribution towards uniform distribution.

$$loss = error(p,\hat{p}) \boxed{-\lambda \sum_i \hat{p}_i log(\hat{p}_i)} \quad \lambda \geq 0,\ \lambda \propto model\ bias,\ \ \lambda \propto \frac{1}{model\ variance}$$

2. Modify data sampling:

- **Data augmentation:** Create more data from available data by randomly cropping, dilating, rotating, adding small amount of noise etc.
- **K-fold Cross-validation:** Divide the data into k groups. Train on (k-1) groups and test on 1 group. Try all k possible combinations.

3. Change training approach:

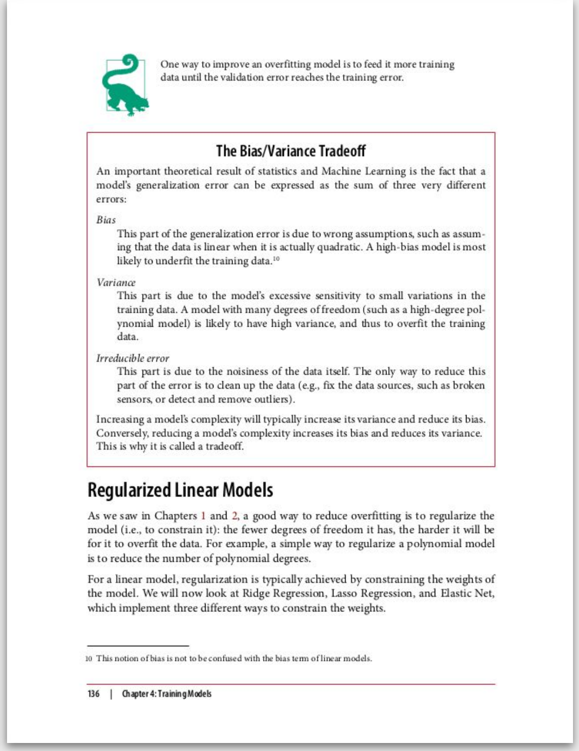
- **Injecting noise:** Add random noise to the weights when they are being learned. It pushes the model to be relatively insensitive to small variations in the weights, hence regularization
- **Dropout:** Generally used for neural networks. Connections between consecutive layers are randomly dropped based on a dropout-ratio and the remaining network is trained in the current iteration. In the next iteration, another set of random connections are dropped.

Figure 2. K-fold CV

Figure 3. Drop-out

Source: https://www.cheatsheets.aqeel-anwar.com

https://sites.google.com/view/datascience-cheat-sheets



One way to improve an overfitting model is to feed it more training data until the validation error reaches the training error.

The Bias/Variance Tradeoff

An important theoretical result of statistics and Machine Learning is the fact that a model's generalization error can be expressed as the sum of three very different errors:

Bias
This part of the generalization error is due to wrong assumptions, such as assuming that the data is linear when it is actually quadratic. A high-bias model is most likely to underfit the training data.[10]

Variance
This part is due to the model's excessive sensitivity to small variations in the training data. A model with many degrees of freedom (such as a high-degree polynomial model) is likely to have high variance, and thus to overfit the training data.

Irreducible error
This part is due to the noisiness of the data itself. The only way to reduce this part of the error is to clean up the data (e.g., fix the data sources, such as broken sensors, or detect and remove outliers).

Increasing a model's complexity will typically increase its variance and reduce its bias. Conversely, reducing a model's complexity increases its bias and reduces its variance. This is why it is called a tradeoff.

Regularized Linear Models

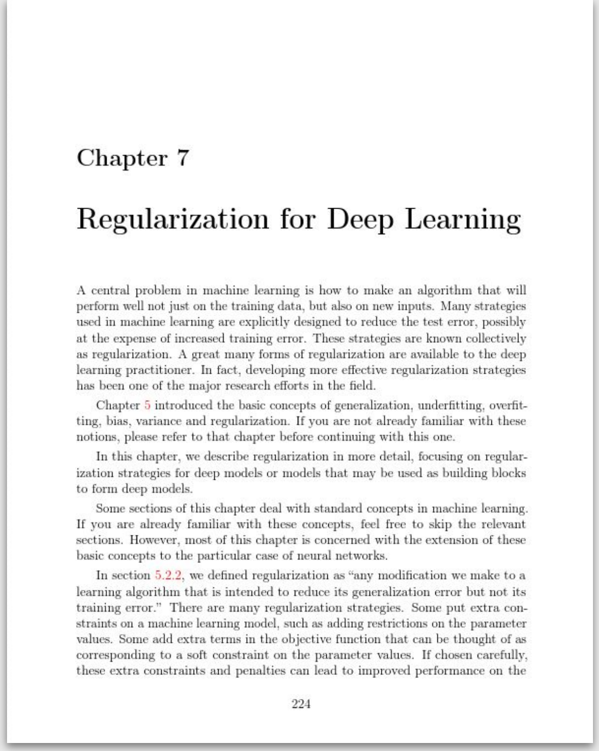
As we saw in Chapters 1 and 2, a good way to reduce overfitting is to regularize the model (i.e., to constrain it): the fewer degrees of freedom it has, the harder it will be for it to overfit the data. For example, a simple way to regularize a polynomial model is to reduce the number of polynomial degrees.

For a linear model, regularization is typically achieved by constraining the weights of the model. We will now look at Ridge Regression, Lasso Regression, and Elastic Net, which implement three different ways to constrain the weights.

[10] This notion of bias is not to be confused with the bias term of linear models.

136 | Chapter 4: Training Models

Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow, Chap. 4



Chapter 7

Regularization for Deep Learning

A central problem in machine learning is how to make an algorithm that will perform well not just on the training data, but also on new inputs. Many strategies used in machine learning are explicitly designed to reduce the test error, possibly at the expense of increased training error. These strategies are known collectively as regularization. A great many forms of regularization are available to the deep learning practitioner. In fact, developing more effective regularization strategies has been one of the major research efforts in the field.

Chapter 5 introduced the basic concepts of generalization, underfitting, overfitting, bias, variance and regularization. If you are not already familiar with these notions, please refer to that chapter before continuing with this one.

In this chapter, we describe regularization in more detail, focusing on regularization strategies for deep models or models that may be used as building blocks to form deep models.

Some sections of this chapter deal with standard concepts in machine learning. If you are already familiar with these concepts, feel free to skip the relevant sections. However, most of this chapter is concerned with the extension of these basic concepts to the particular case of neural networks.

In section 5.2.2, we defined regularization as "any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error." There are many regularization strategies. Some put extra constraints on a machine learning model, such as adding restrictions on the parameter values. Some add extra terms in the objective function that can be thought of as corresponding to a soft constraint on the parameter values. If chosen carefully, these extra constraints and penalties can lead to improved performance on the

224

https://www.deeplearningbook.org/contents/regularization.html

λ lambda

2. Entre qual intervalo de valores tem que estar o hiperparâmetro lambda da regularização?
    - If your **lambda value is too high**, your model will be simple, but you run the risk of **underfitting** your data.
    - If your **lambda value is too low**, your model will be more complex, and you run the risk of **overfitting** your data.
    - The ideal value of lambda produces a model that generalizes well to new, previously unseen data. Unfortunately, that ideal value of **lambda is data-dependent**, so you'll need to do some tuning.
3. Se eu tiver rodando as épocas eu devo atualizar o lambda junto com os outros parâmetros? Ou o lambda é o mesmo para todo o processo?

https://neptune.ai/blog/fighting-overfitting-with-l1-or-l2-regularization

Bias/Variance Trade-off

4. Fiquei confuso entre os conceitos de bias e variância. Não é apenas uma nova nomenclatura para underfitting e overfitting ou tem algum detalhe a mais que eu não peguei?
    - The **bias error** is an error from erroneous assumptions in the **learning algorithm**. **High bias** can cause an algorithm to miss the relevant relations between features and target outputs (**underfitting**).
    - The **variance** is an error from sensitivity to small fluctuations **in the training set**. **High variance** may result from an algorithm modeling the random **noise** in the training data (**overfitting**).
5. Não ficou claro para mim, o que é “erro irredutível”? É um erro que não podemos evitar em nosso modelo? Tem uma causa que provoca esse erro?

Overfitting

6. Quando tenho um conjunto de treino com, por exemplo, 98% de acerto, um conjunto de validação com 97% e um conjunto de teste que resultou em 60%, posso dizer também que houve overfitting ou seria overfitting somente se o 60% fosse do conjunto de validação e a gente nem olha para o teste nessa nomenclatura?

7. Existe alguma situação que fazer a regularização não é aconselhável?

Maior Dúvida da Aula

[Machine Learning] Datasets

MC886/MO444, September 15, 2022

1. Não sabia da quantidade de fontes diferentes de dados, pois toda vida só trabalhei com dados privados. Onde se pode procurar dados locais de serviços públicos do estado de São Paulo?
2. Posso montar o meu próprio dataset para ser usado no projeto final?
3. Deveremos já ter enviado a proposta de trabalho para o Trabalho Final e o conjunto de dados com que iremos trabalhar até o dia 28/09?
4. Temos que usar uma dessas bases de dados? Podemos criar uma ou procurar por outras?
5. Olá professora. Existe algum novo dataset que está sendo construído com o intuito de 'substituir' o Imagenet ? principalmente levando em consideração os possíveis vieses presentes no mesmo. Ou a ideia seria atualizar o Imagenet e realizar uma curadoria no mesmo?
6. Não entendi como determinar se um dataset é muito artificial. Seria por algum tipo de análise exploratória dos dados?

7. É possível utilizar um dataset privado* para o projeto final? *São dados estruturados de uma empresa.
8. Quais são os pontos mais importantes para se analisar primordialmente se a base de dados é boa o suficiente para o nosso problema?
9. Eu já tive um trabalho onde eu precisava manipular dados, no caso eu usei o Kaggle, porém naquela matéria eu usei dois bancos de dados de diferentes fontes e fiz específico para o meu trabalho. No trabalho final, posso fazer algo parecido?
10. Eu gostaria de saber quais bibliotecas vamos utilizar no projeto, estou com um pouco de ansiedade pois não sei se tenho o domínio de todas as ferramentas que serão necessárias para desenvolver o projeto.
11. Na aula foi apresentado a evolução dos desempenhos das técnicas de aprendizado de máquina no horizonte de alguns anos. Minha pergunta é: Hoje, qual seria o próximo grande avanço almejado para as técnicas de aprendizado de máquina?

Testing and Error Metrics

Machine Learning

(Largely based on slides from Luis Serrano)

Prof. Sandra Avila

Institute of Computing (IC/Unicamp)

MC886/MO444, September 15, 2022

How well is my model doing?

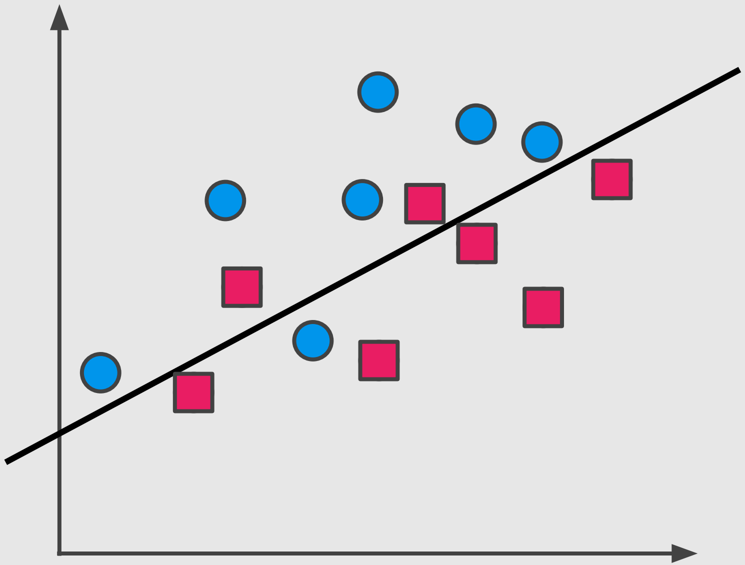
"We say that a machine learns with respect to a particular task T, performance metric P, and type of experience E, if the systems reliably improves its performance P at task T, following experience E."

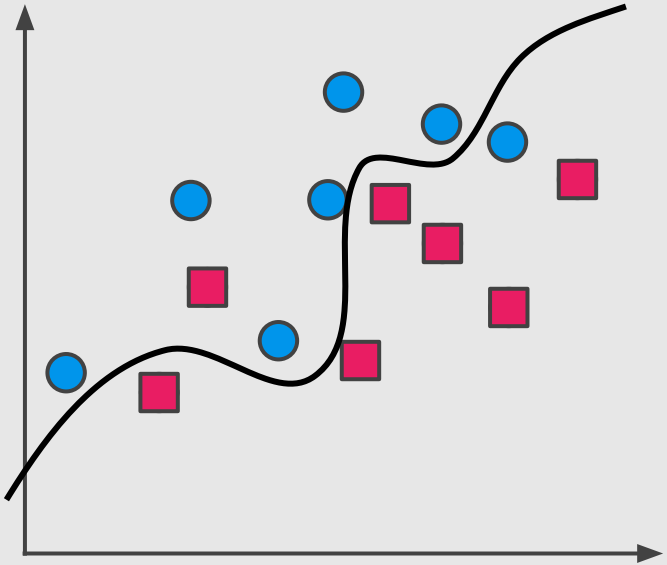
[Tom M. Mitchell, 1997]

Today’s Agenda

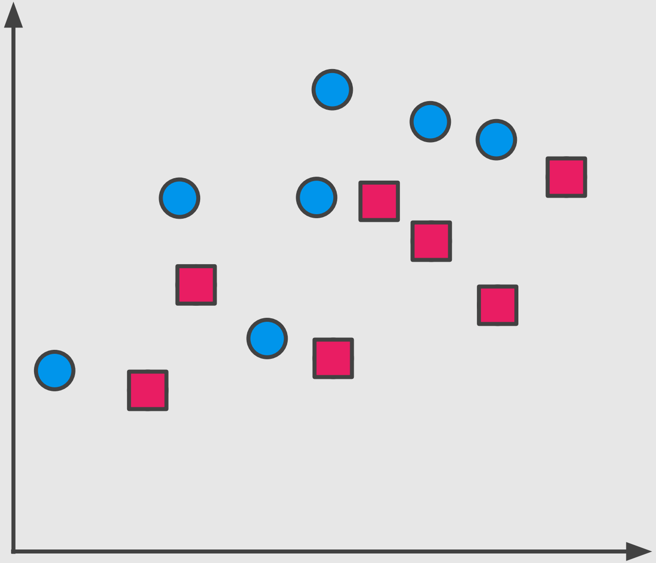
- Testing and Error Metrics
    - Training, Testing
    - Accuracy
    - Precision
    - Recall
    - F-Score

Which model is better?





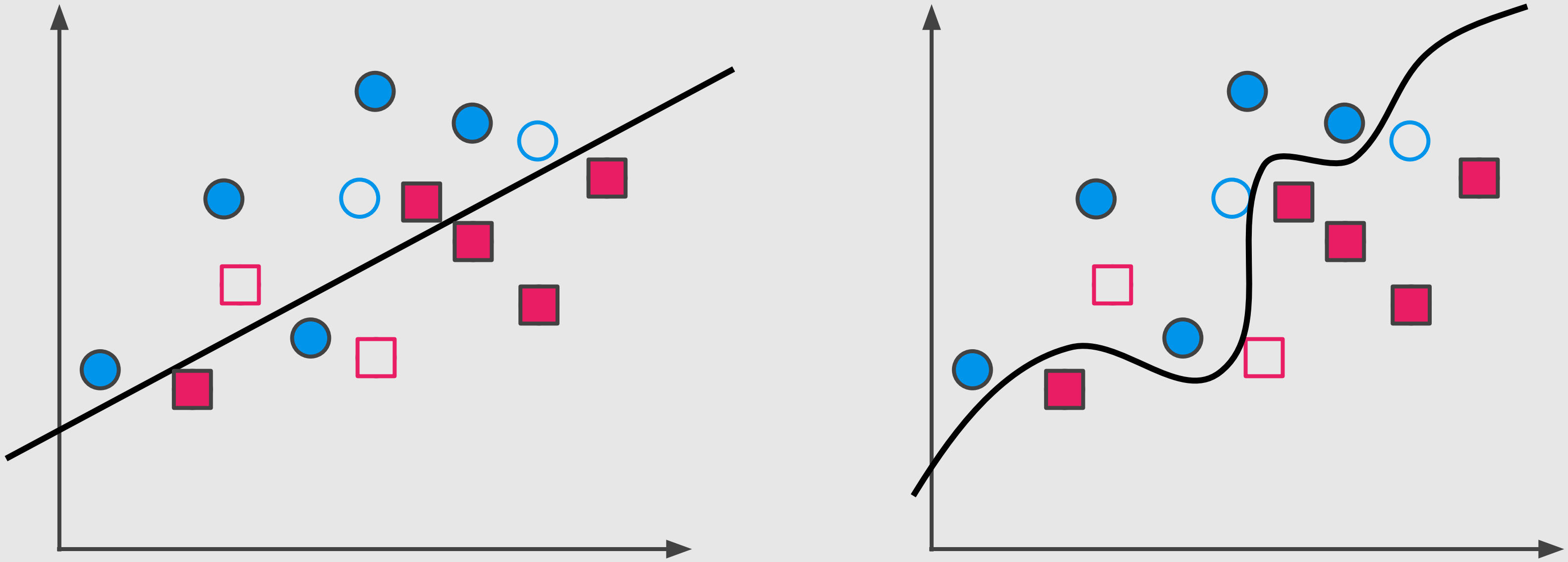
Why validating?



● ■ Training

○ □ Validation

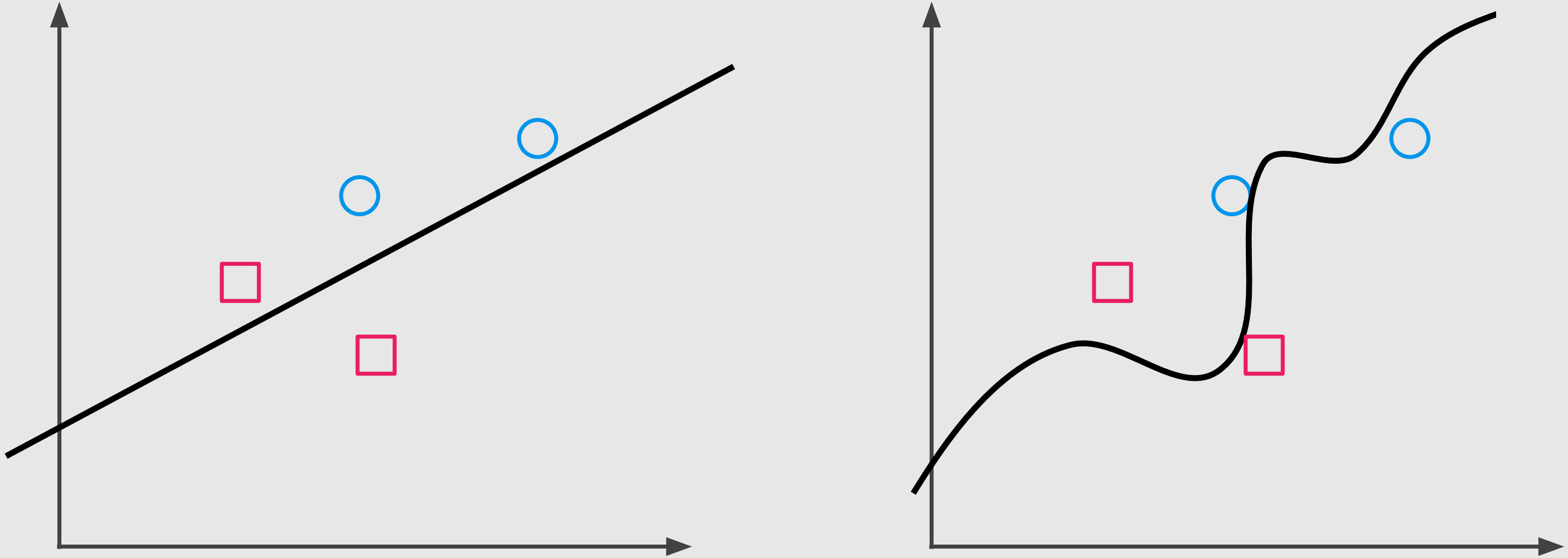
Why validating?



Why validating?

● ■ Training

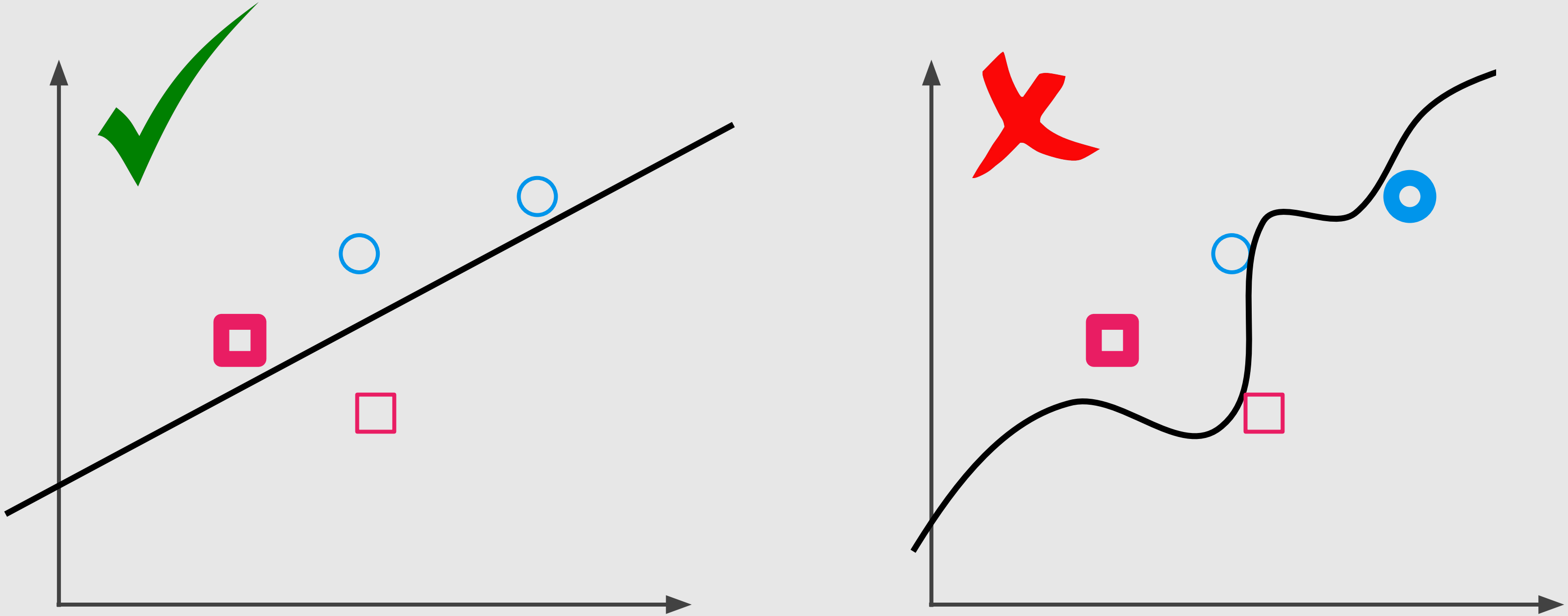
○ □ Validation



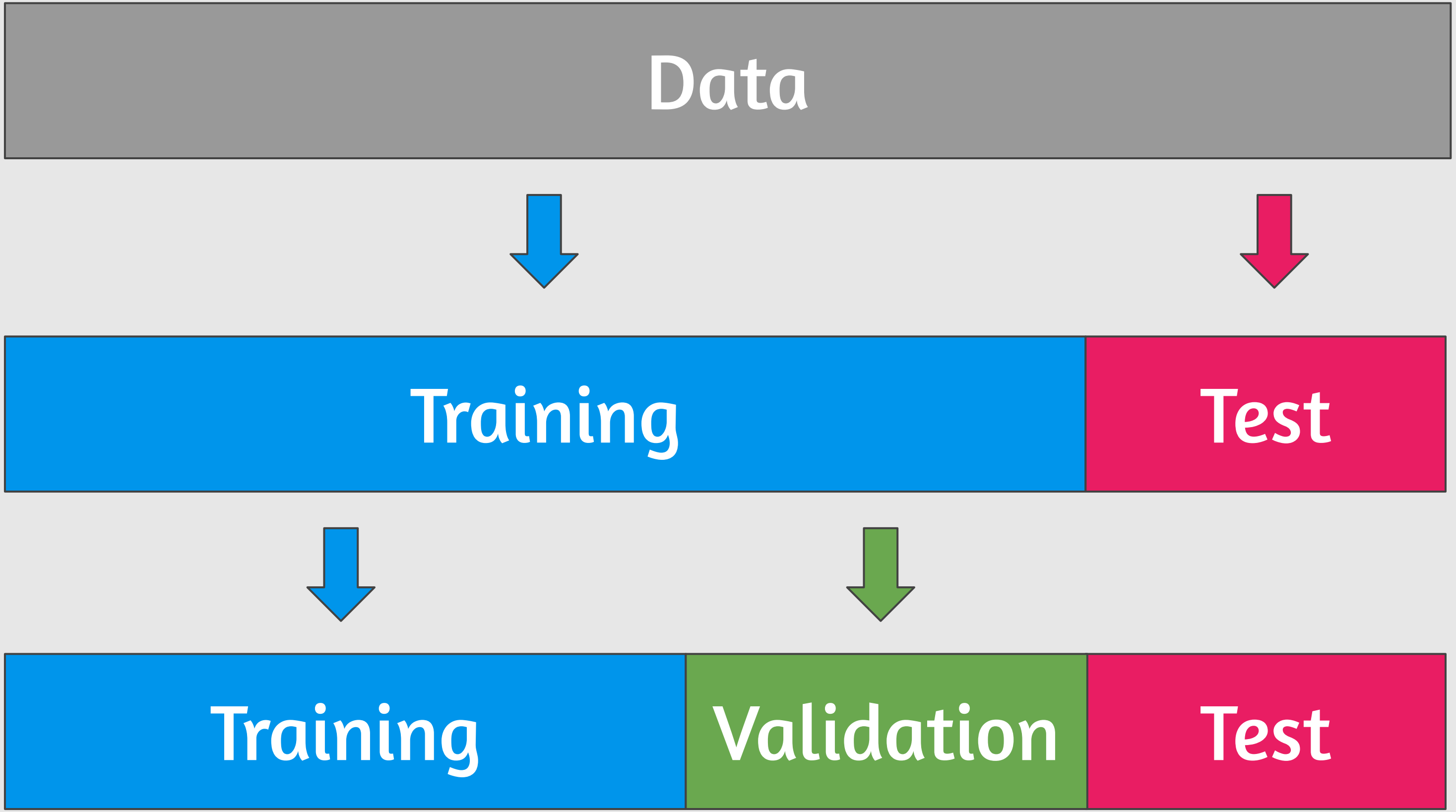
● ■ Training

○ □ Validation

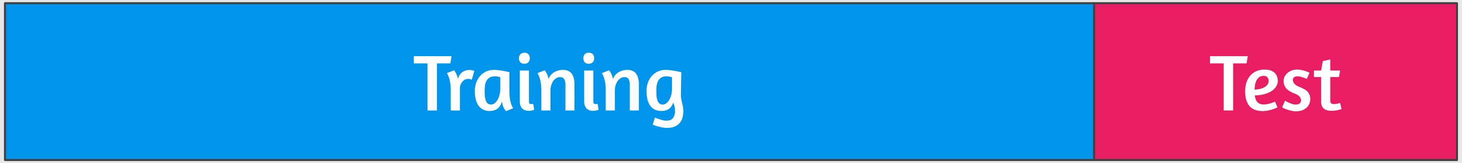
Why validating?



Friends don’t let friends use testing data for training


Data
Training
Test
Training
Validation
Test

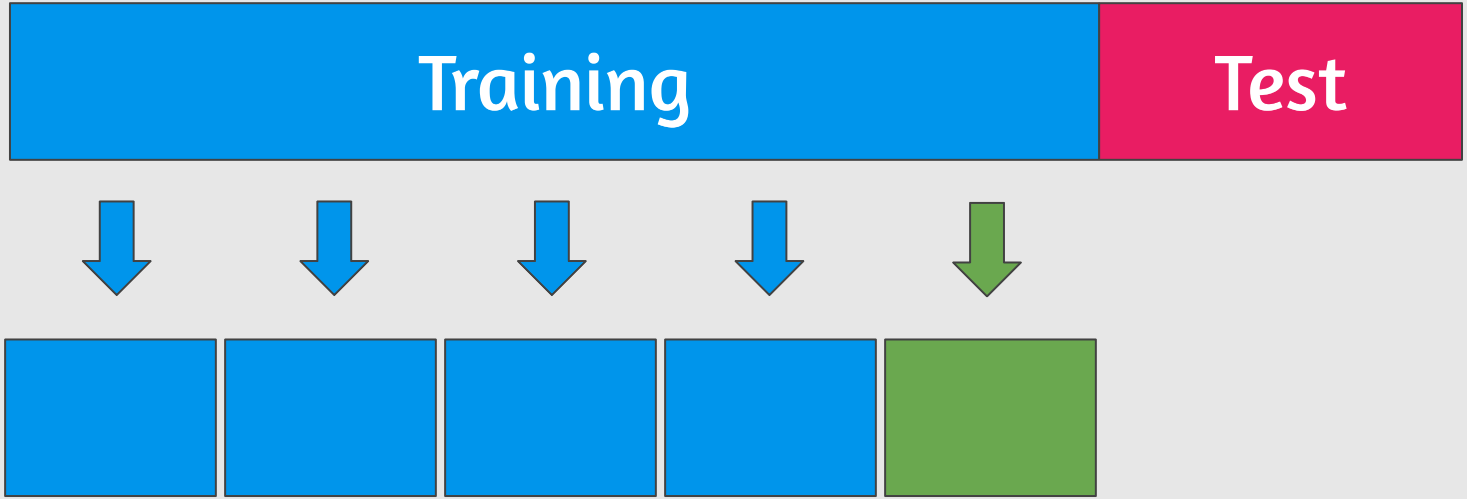
How do we not ‘lose’ the training data?



Training

Test

k-fold Cross Validation

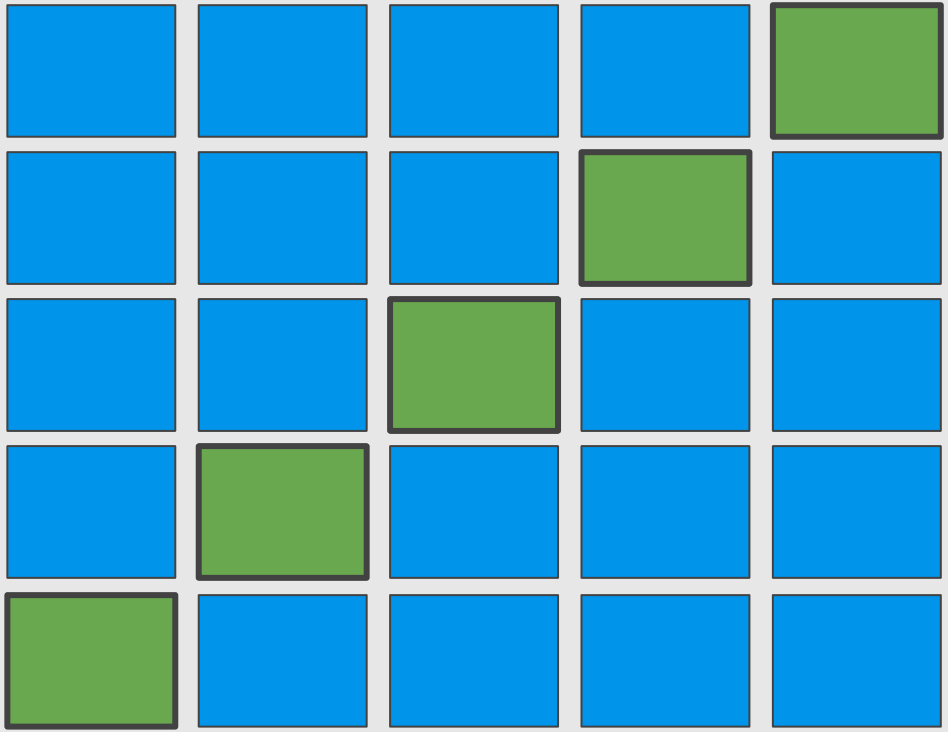


k-fold Cross Validation

🟦 Training

🟩 Validation

k = 5



k×2-fold Cross Validation

🟦 Training

🟩 Validation

k = 5

k×2-fold Cross Validation

🟦 Training

🟩 Validation

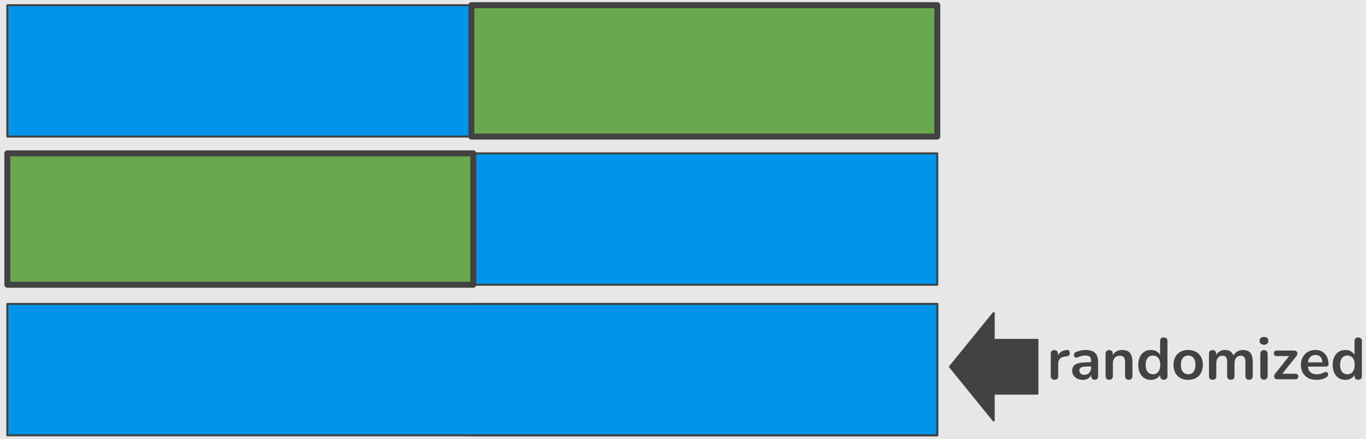
k = 5

randomized

k×2-fold Cross Validation

🟦 Training
🟩 Validation

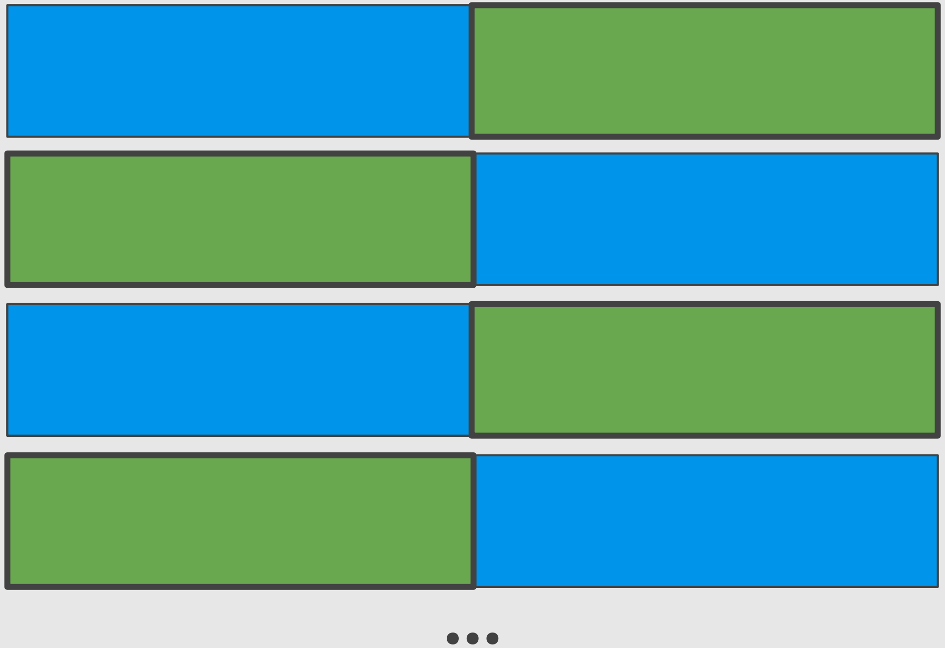
k = 5



k×2-fold Cross Validation

Training
Validation

k = 5

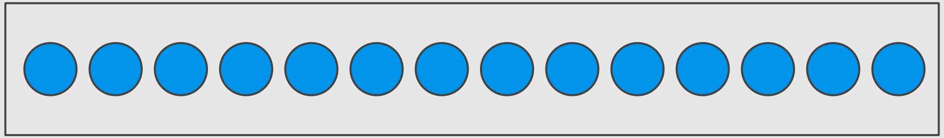


k times = k×2 folds

Randomizing in Cross Validation

● Training

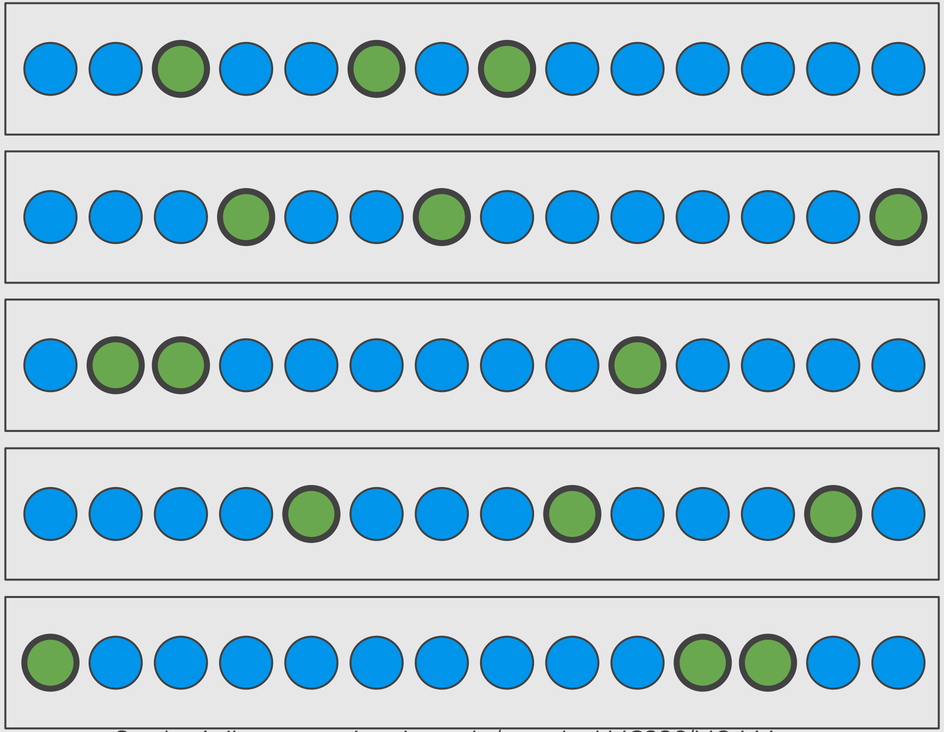
● Validation



Randomizing in Cross Validation

● Training

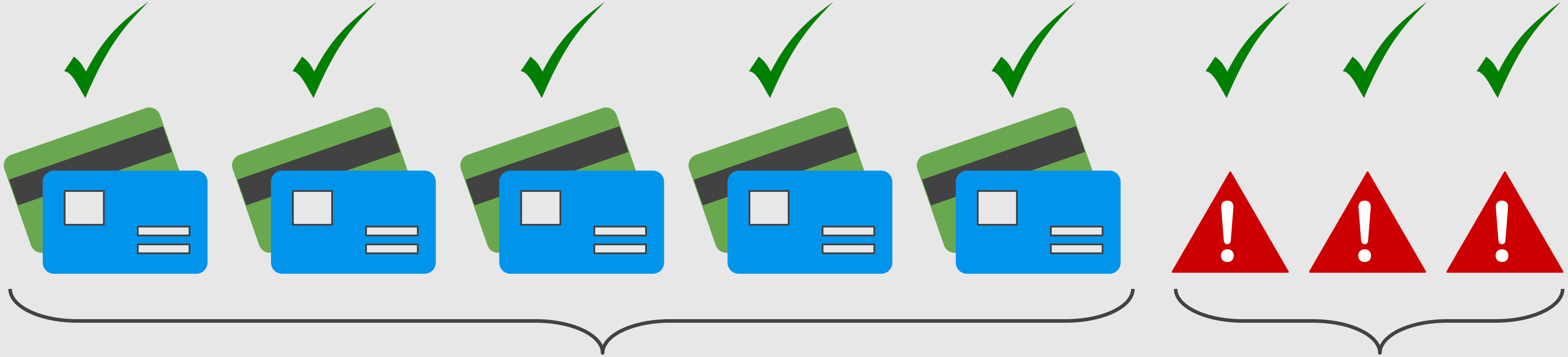
● Validation



Evaluation Metrics

How well is my model doing?

Credit Card Fraud



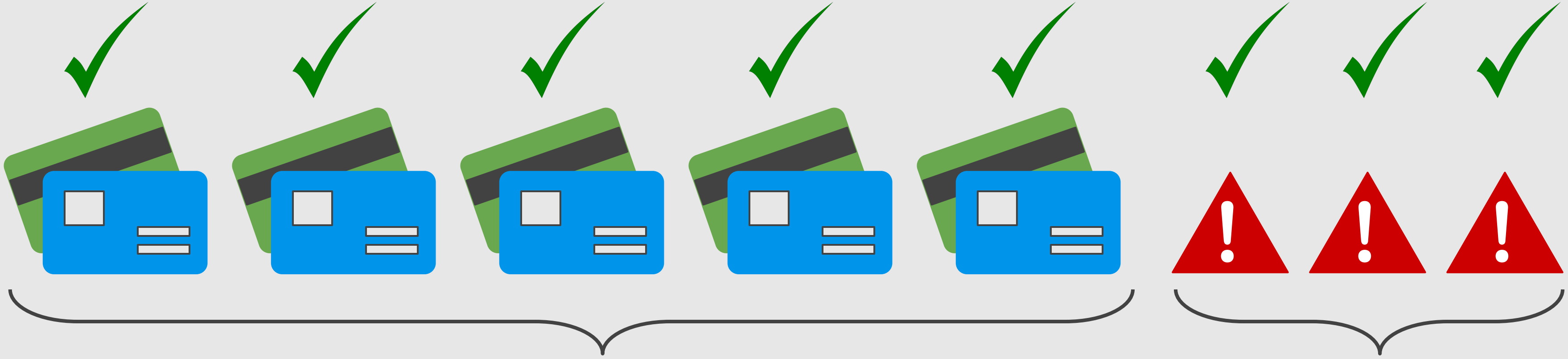
284,335

472

Model: All transactions are good.

$$\text{Correct} = \frac{284{,}335}{284{,}807} = 99.83\%$$

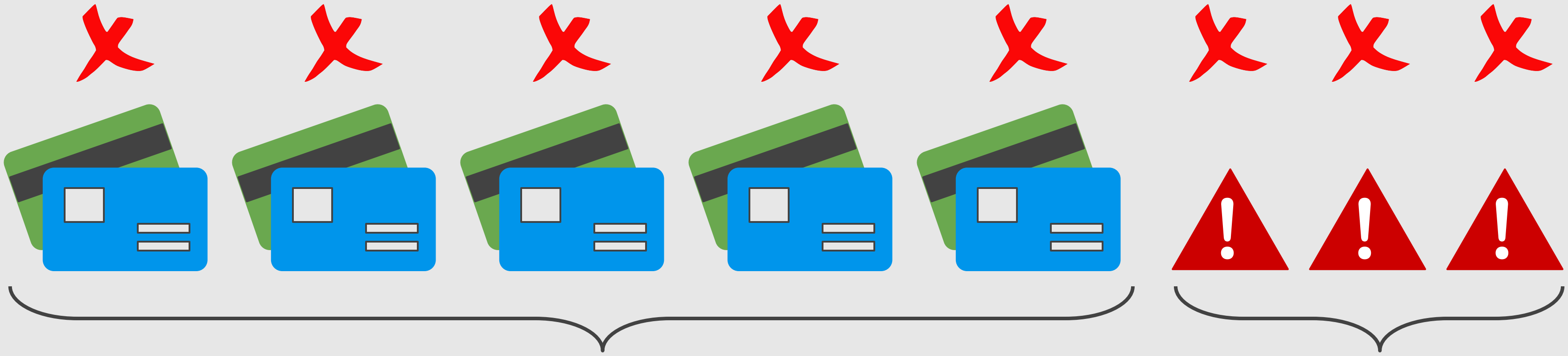
Credit Card Fraud



Model: All transactions are good.

Problem: I’m not catching any of the bad ones!

Credit Card Fraud



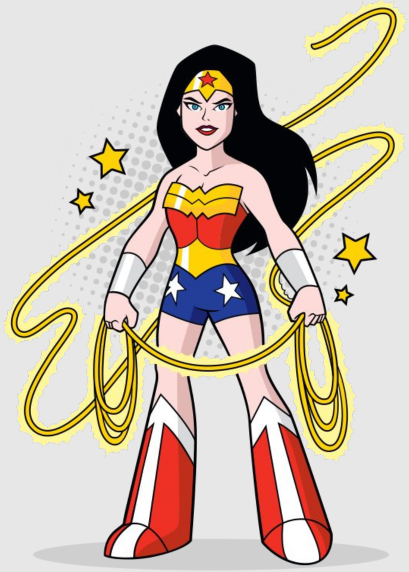
284,335

472

Model: All transactions are fraudulent.

Problem: I’m accidently catching all the good ones!

Medical Model



Health

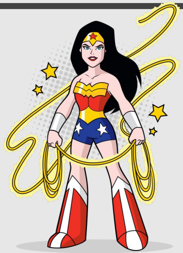
Sick

Spam Classifier Model

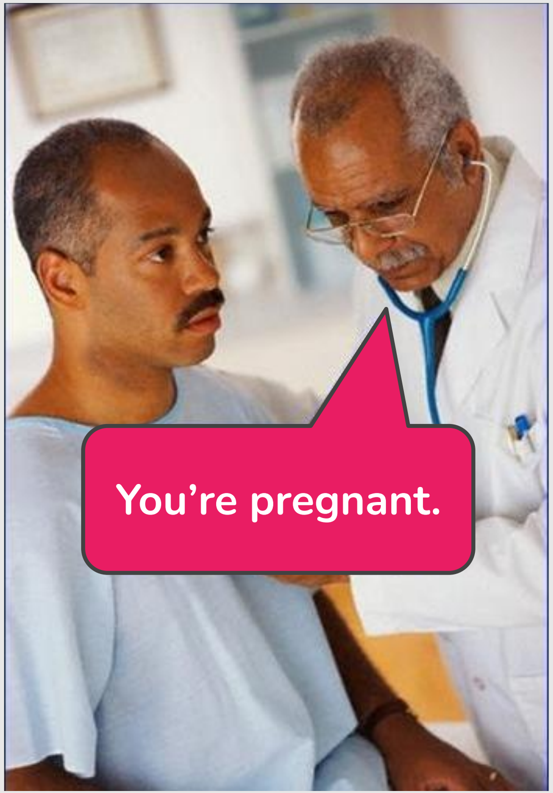
Not Spam

Spam

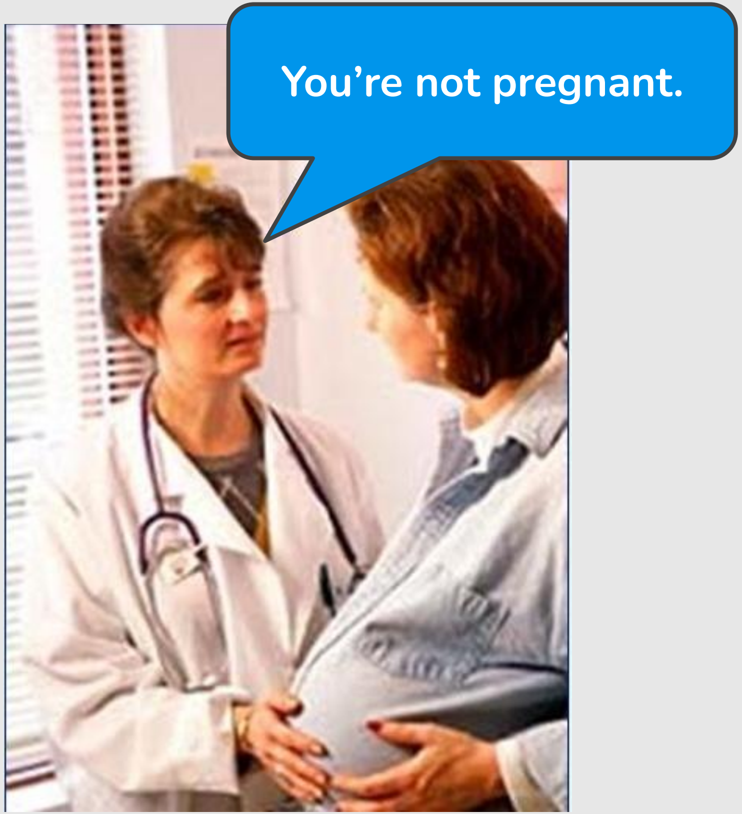
Confusion Matrix Table

	Diagnosed Sick	Diagnosed Healthy
Sick	True Positive	False Negative
Healthy	False Positive 	True Negative

Type I Error
(False Positive)



Type II Error
(False Negative)



Confusion Matrix Table

10,000
patients

Diagnosis

Patients

	Diagnosed Sick	Diagnosed Healthy
Sick	1000	200
Healthy	800	8000

Confusion Matrix Table

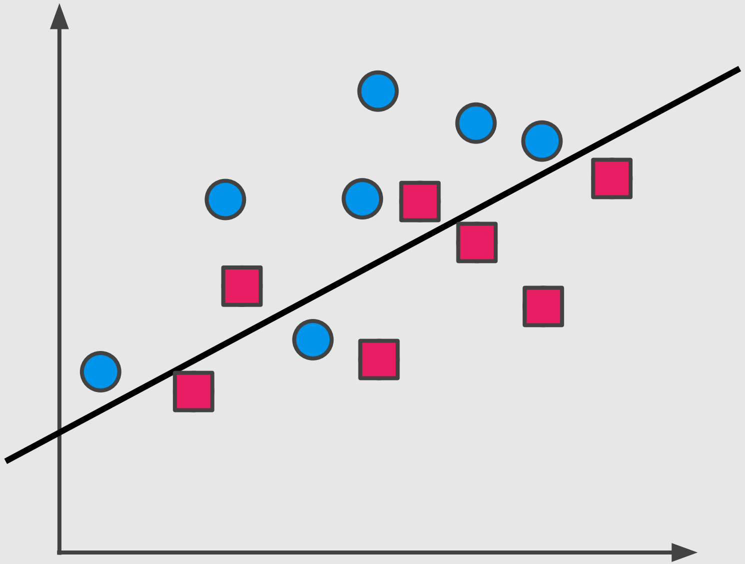
	Sent to Spam Folder	Sent to Inbox
Spam	True Positive	False Negative
Not Spam	False Positive	True Negative

Confusion Matrix Table

1,000 emails

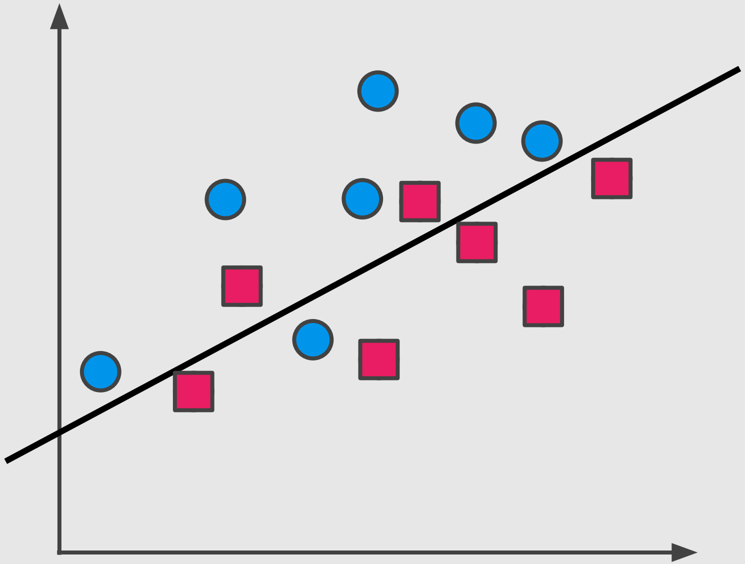
		Folder	
		Spam Folder	Inbox
Email	Spam	100	170
	Not Spam	30	700

Confusion Matrix Table



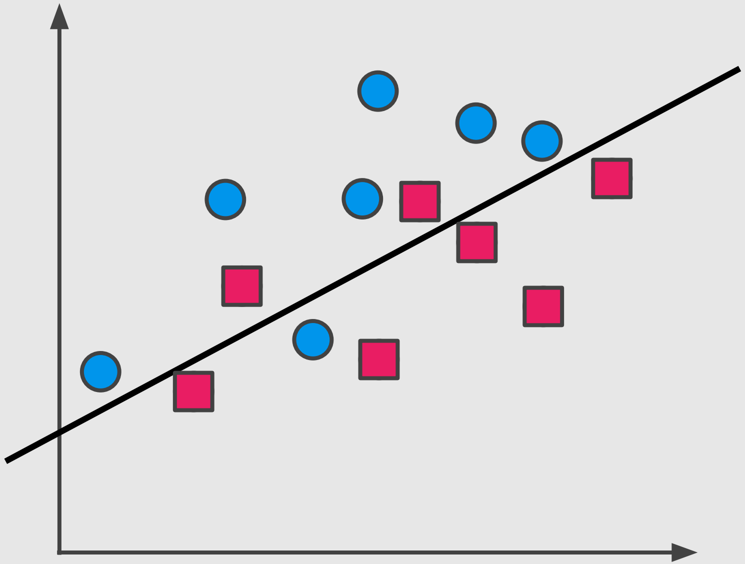
		Prediction	
		Guessed Positive	Guessed Negative
Data	Positive		
	Negative		

Confusion Matrix Table



	Prediction	
Data	Guessed Positive	Guessed Negative
Positive	6 True positives	
Negative		

Confusion Matrix Table

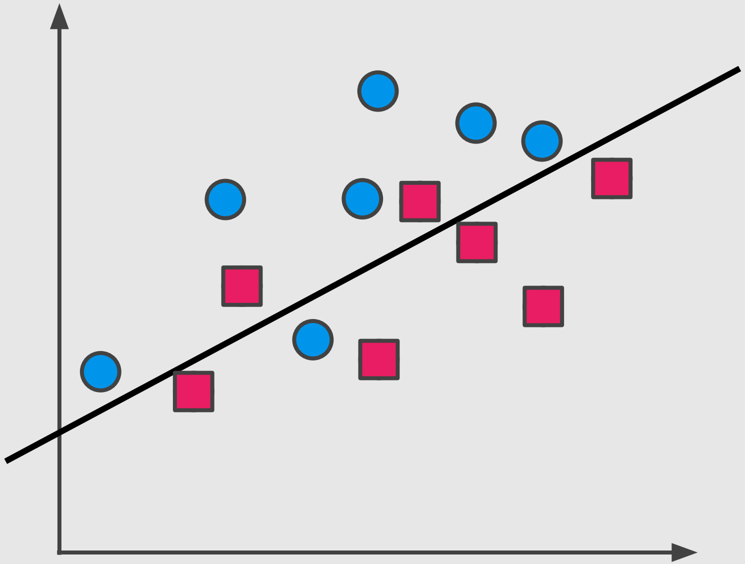


Prediction

Data

	Guessed Positive	Guessed Negative
Positive	6 True positives	
Negative		5 True negatives

Confusion Matrix Table

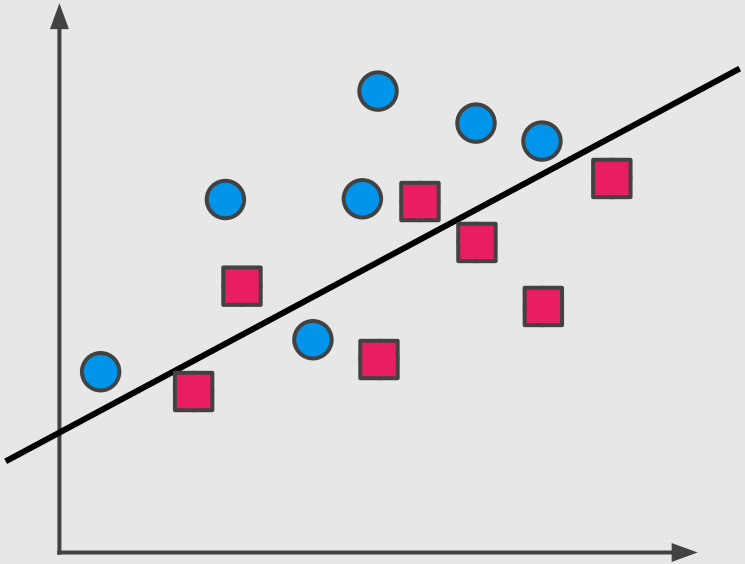


Prediction

Data

	Guessed Positive	Guessed Negative
Positive	6 True positives	1 False negative
Negative		5 True negatives

Confusion Matrix Table



Prediction

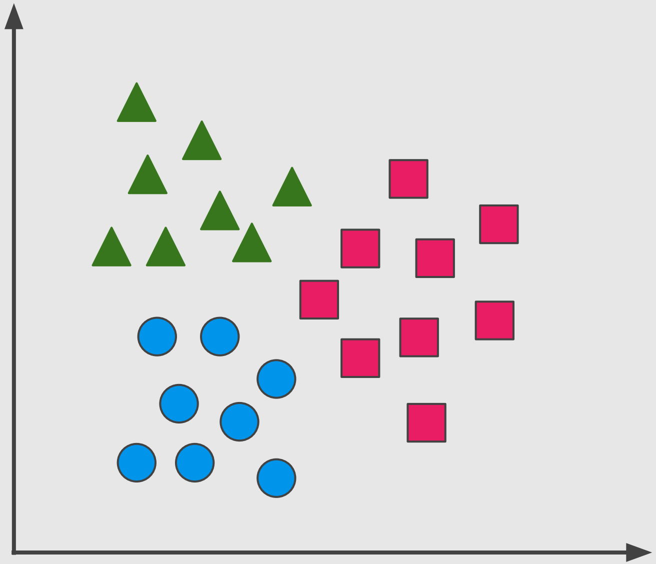
Data	Guessed Positive	Guessed Negative
Positive	6 True positives	1 False negative
Negative	2 False positives	5 True negatives

Confusion Matrix Table (n classes)

Class 1: ▲

Class 2: ■

Class 3: ●

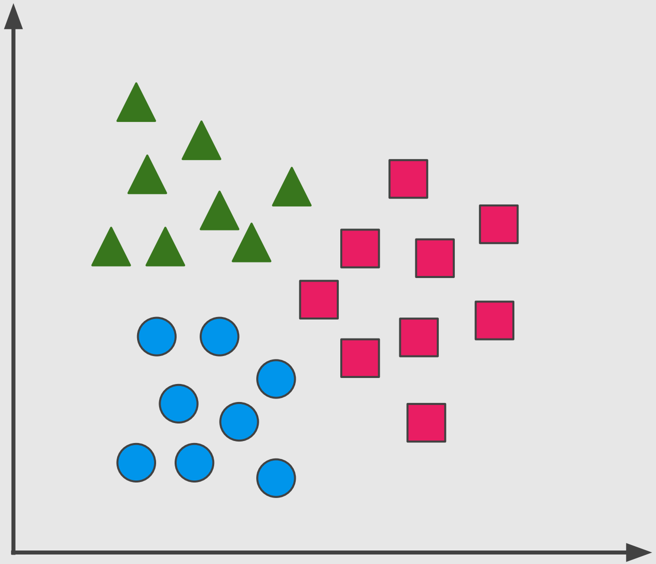


Confusion Matrix Table (n classes)

Class 1: ▲

Class 2: ■

Class 3: ●



Predicted Class

True Class

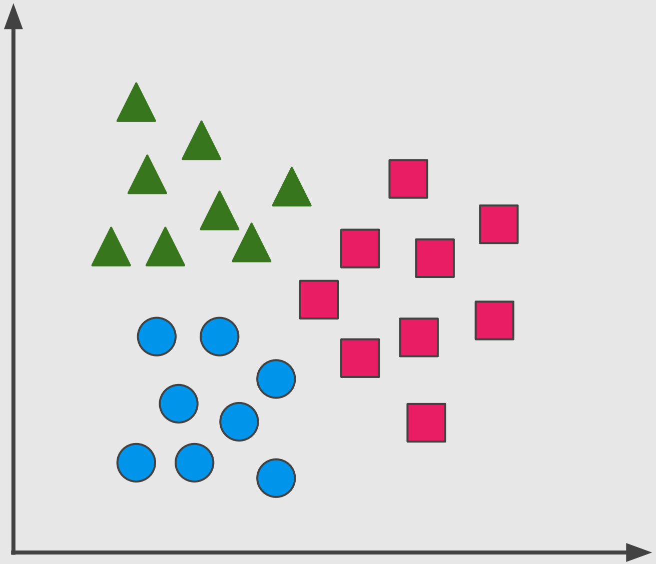
	Guessed Class 1	Guessed Class 2	Guessed Class 3
Class 1			
Class 2			
Class 3			

Confusion Matrix Table (n classes)

Class 1: ▲

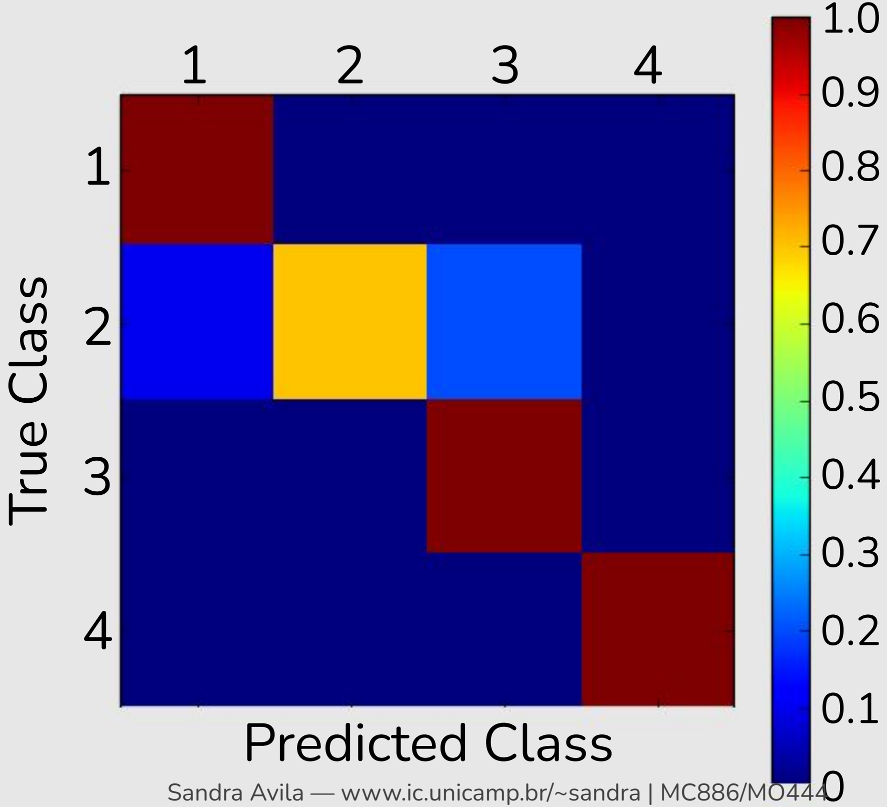
Class 2: ■

Class 3: ●

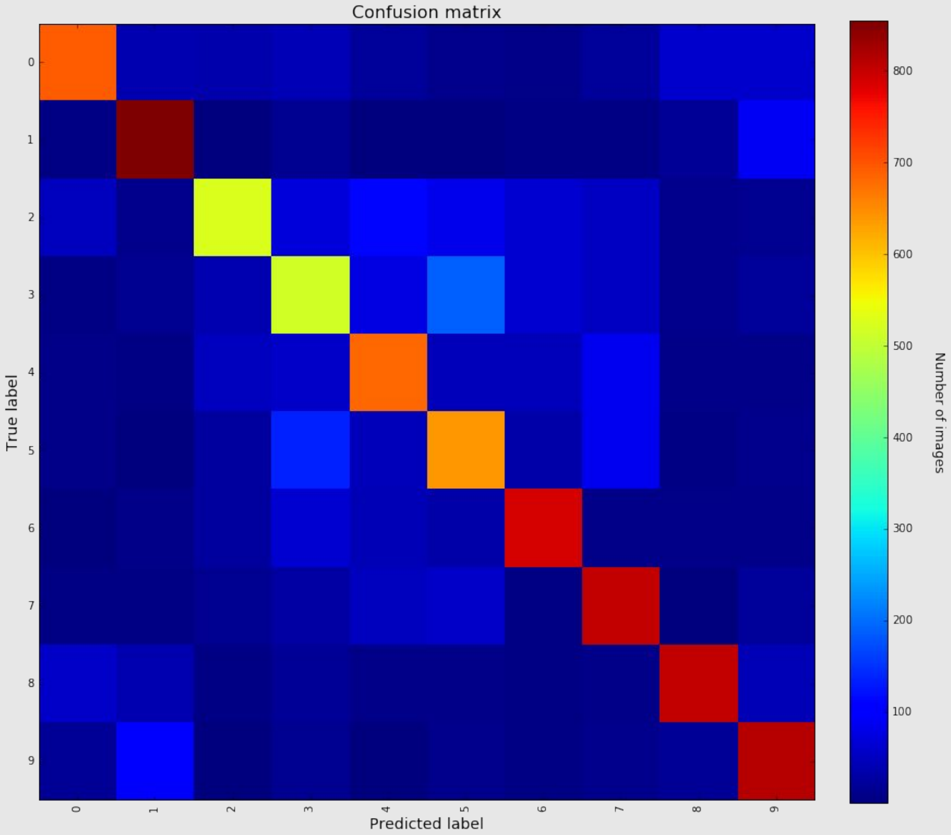


		Predicted Class		
		Guessed Class 1	Guessed Class 2	Guessed Class 3
True Class	Class 1	5	2	1
	Class 2	3	6	0
	Class 3	0	1	7

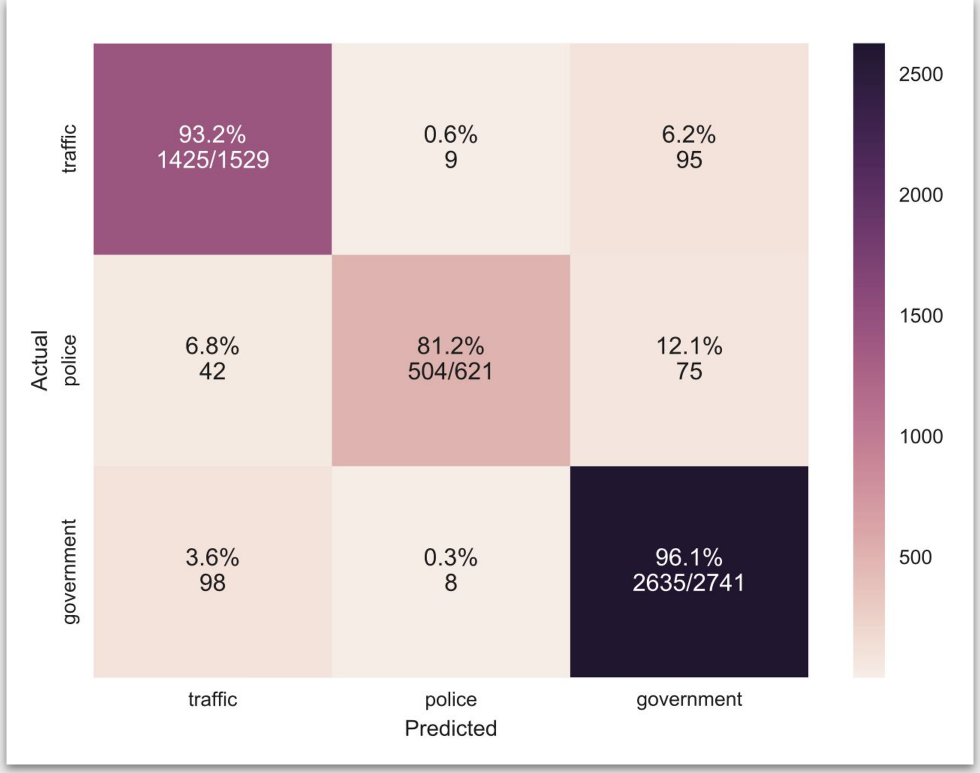
Confusion Matrix Table (*n* classes)



Confusion Matrix Table (*n* classes)



Confusion Matrix Table (n classes)



Actual \ Predicted	traffic	police	government
traffic	93.2% 1425/1529	0.6% 9	6.2% 95
police	6.8% 42	81.2% 504/621	12.1% 75
government	3.6% 98	0.3% 8	96.1% 2635/2741

https://gist.github.com/hitvoice/36cf44689065ca9b927431546381a3f7#file-plot_confusion_matrix-py

Today’s Agenda

- Testing and Error Metrics
    - Training, Testing
    - **Accuracy**
    - Precision
    - Recall
    - F-Score

Accuracy

Diagnosis

Patients

	Diagnosed Sick	Diagnosed Healthy
Sick	1,000	200
Healthy	800	8,000

Accuracy

		Diagnosis	
		Diagnosed Sick	Diagnosed Healthy
Patients	Sick	1,000	200
	Healthy	800	8,000

Accuracy:
Out of all the **patients**, how many did we classify correctly?

Accuracy

		Diagnosis	
		Diagnosed Sick	Diagnosed Healthy
Patients	Sick	1,000	200
	Healthy	800	8,000

Accuracy:
Out of all the **patients**, how many did we classify correctly?

$$\text{Accuracy} = \frac{1{,}000 + 8{,}000}{\quad}$$

Accuracy

		Diagnosis	
		Diagnosed Sick	Diagnosed Healthy
Patients	Sick	1,000	200
	Healthy	800	8,000

Accuracy:
Out of all the **patients**, how many did we classify correctly?

$$\text{Accuracy} = \frac{1{,}000 + 8{,}000}{10{,}000} = 90\%$$

Accuracy

		Folder	
		Spam Folder	Inbox
Email	Spam	100	170
	Not Spam	30	700

Accuracy:
Out of all the **emails**, how many did we classify correctly?

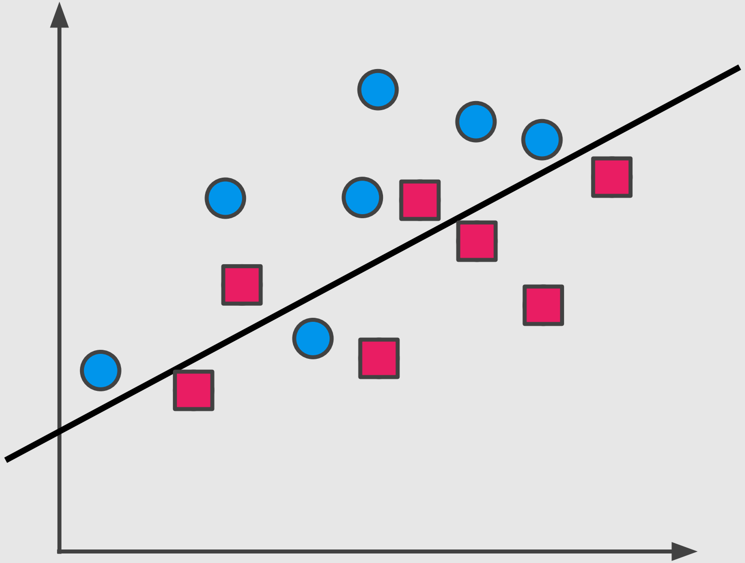
Accuracy

		Folder	
		Spam Folder	Inbox
Email	Spam	100	170
	Not Spam	30	700

Accuracy:
Out of all the **emails**, how many did we classify correctly?

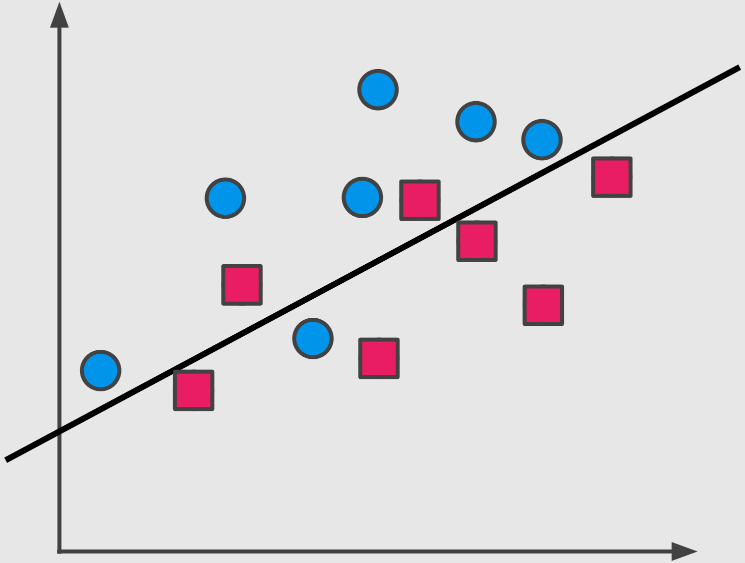
$$\text{Accuracy} = \frac{100 + 700}{1{,}000} = 80\%$$

Accuracy



Accuracy:
Out of all the **data**, how many points did we classify correctly?

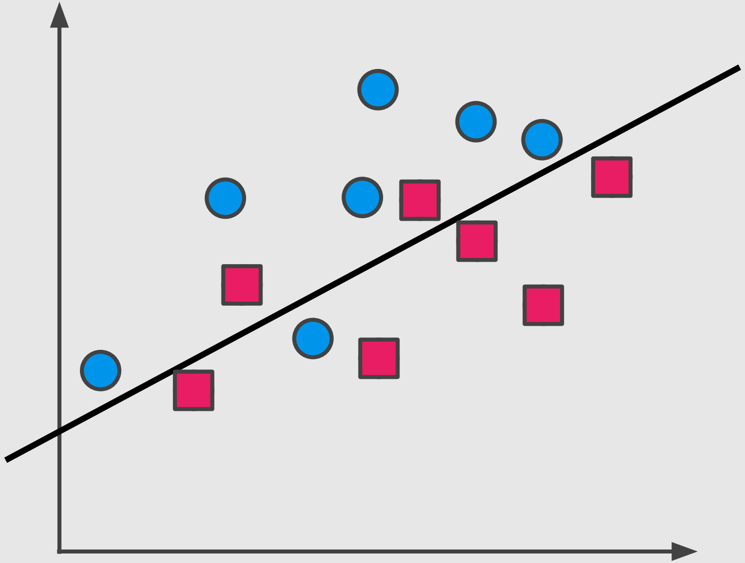
Accuracy



Accuracy:
Out of all the **data**, how many points did we classify correctly?

$$\text{Accuracy} = \frac{\text{Correctly Classified Points}}{\text{All points}}$$

Accuracy



Accuracy:
Out of all the **data**, how many points did we classify correctly?

Accuracy =

$$\frac{\text{Correctly Classified Points}}{\text{All points}}$$

$$\frac{11}{11 + 3} = 78.57\%$$

Accuracy

		Prediction	
		Fraudulent	Not Fraudulent
Transactions	Fraudulent	0	472
	Not Fraudulent	0	284,335

Accuracy:
Out of all the **transactions**, how many did we classify correctly?

$$\text{Accuracy} = \frac{0 + 284{,}335}{284{,}807} = 99.83\%$$

Overall (Normalized) Accuracy

		Prediction	
		Fraudulent	Not Fraudulent
Transactions	Fraudulent	0	472
	Not Fraudulent	0	284,335

Overall Accuracy =

$$\frac{\frac{\text{TP}}{\text{TP + FN}} + \frac{\text{TN}}{\text{TN + FP}}}{2} =$$

$$\frac{\frac{0}{0 + 472} + \frac{284{,}335}{284{,}335 + 0}}{2} =$$

$$\frac{0 + 100}{2} = 50\%$$

Overall (Normalized) Accuracy

Accuracy = 80%

		Folder	
		Spam Folder	**Inbox**
Email	**Spam**	100	170
	Not Spam	30	700

Overall Accuracy =

$$\frac{\frac{TP}{TP + FN} + \frac{TN}{TN + FP}}{2} =$$

$$\frac{\frac{100}{100 + 170} + \frac{700}{700 + 30}}{2} =$$

$$\frac{37.0 + 95.9}{2} = 66.5\%$$

Overall (Normalized) Accuracy

Diagnosis

Patients

	Diagnosed Sick	Diagnosed Healthy
Sick	1,000	200
Healthy	800	8,000

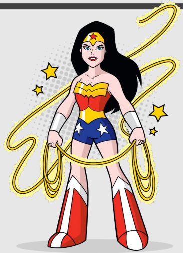
Accuracy = 90%

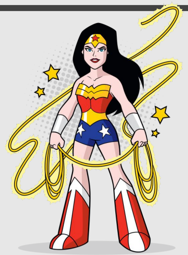
Overall Accuracy =

$$\frac{\frac{TP}{TP + FN} + \frac{TN}{TN + FP}}{2} =$$

$$\frac{\frac{1000}{1000 + 200} + \frac{8000}{8000 + 800}}{2} =$$

$$\frac{83.3 + 90.9}{2} = 87.1\%$$

	Diagnosed Sick	Diagnosed Healthy
Sick	True Positive	False Negative
Healthy	False Positive 	True Negative

	Diagnosed Sick	Diagnosed Healthy
Sick		False Negative
Healthy	False Positive 	

	Sent to Spam Folder	Sent to Inbox
Spam	True Positive ☠	False Negative ☠
Not Spam	False Positive 👵	True Negative 👵

	Sent to Spam Folder	Sent to Inbox
Spam		False Negative
Not Spam	False Positive	

Evaluation Metrics

Medical Model

False positives ok
False negatives **NOT** ok

Spam Detector

False positives **NOT** ok
False negatives ok

Evaluation Metrics

Medical Model

False positives ok
False negatives **NOT** ok

High Recall

Spam Detector

False positives **NOT** ok
False negatives ok

High Precision

Today's Agenda

- Testing and Error Metrics
    - Training, Testing
    - Accuracy
    - **Precision**
    - Recall
    - F-Score

Precision

		Diagnosis	
		Diagnosed Sick	Diagnosed Healthy
Patients	Sick	1,000	200
	Healthy	800	8,000

Precision

		Diagnosis	
		Diagnosed Sick	Diagnosed Healthy
Patients	Sick	1,000	200
	Healthy	800	8,000

Precision:
Out of all the patients we diagnosed with illness, how many were actually sick?

Precision

		Diagnosis	
		Diagnosed Sick	Diagnosed Healthy
Patients	Sick	1,000	200
	Healthy	800	8,000

Precision:
Out of all the patients we diagnosed with illness, how many were actually sick?

Precision

Diagnosis

		Diagnosed Sick	Diagnosed Healthy
Patients	Sick	1,000	200
	Healthy	800	8,000

Precision:
Out of all the patients we diagnosed with illness, how many were actually sick?

$$\text{Precision} = \frac{1{,}000}{1{,}000 + 800} = 55.7\%$$

Precision

		Folder	
		Spam Folder	Inbox
Email	Spam	100	170
	Not Spam	30	700

Precision:
Out of all the emails sent to the spam inbox, how many did were actually spam?

Precision

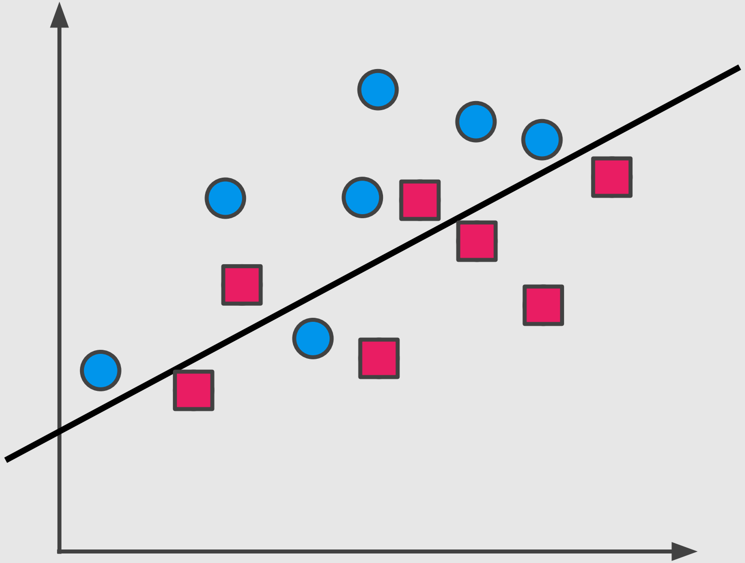
		Folder	
		Spam Folder	Inbox
Email	Spam	100	170
	Not Spam	30	700

Precision:
Out of all the emails sent to the spam inbox, how many did were actually spam?

Precision =

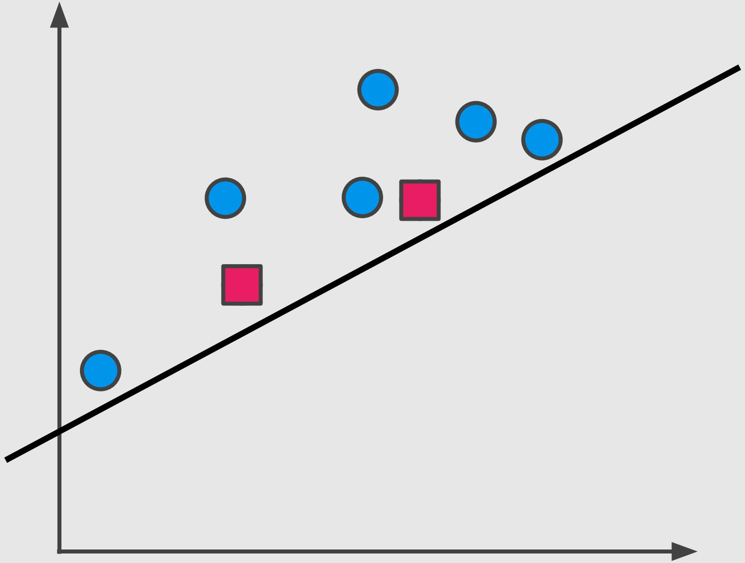
$$\frac{100}{100 + 300} = 76.9\%$$

Precision



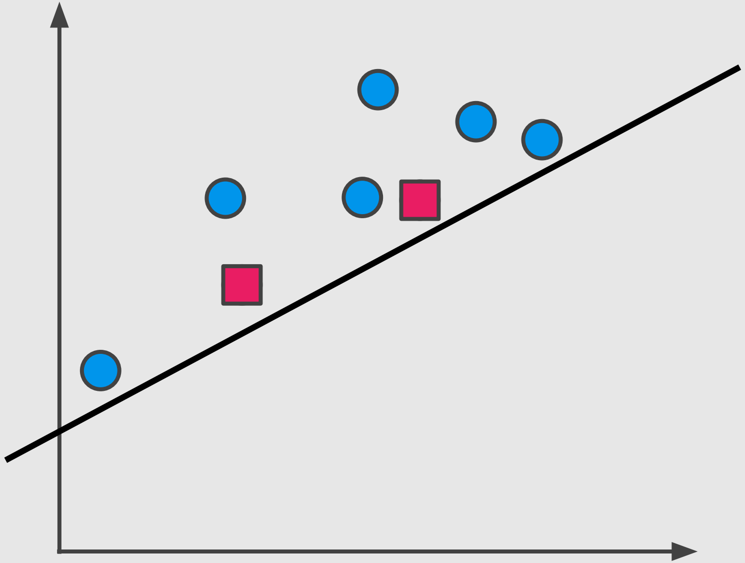
Precision:
Out of all the points we've predicted to be positive, how many are correct?

Precision



Precision:
Out of all the points we've predicted to be positive, how many are correct?

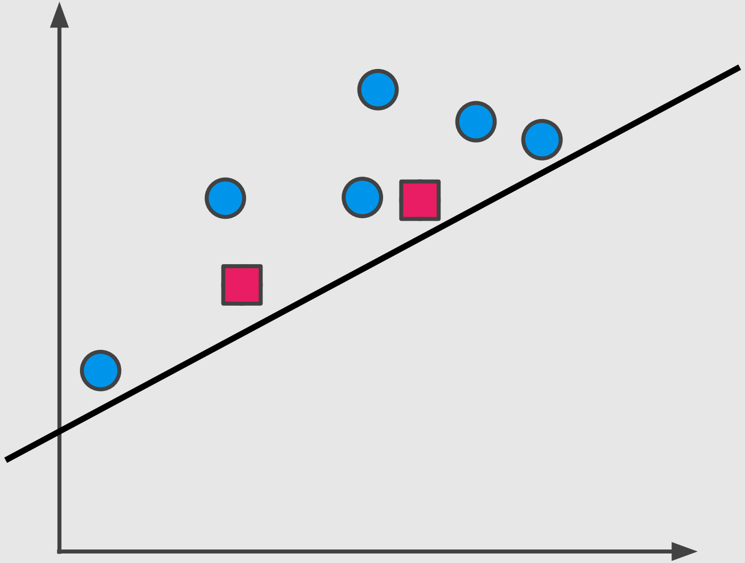
Precision



Precision:
Out of all the points we've predicted to be positive, how many are correct?

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

Precision



Precision:
Out of all the points we've predicted to be positive, how many are correct?

Precision =

$$\frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

$$\frac{6}{6 + 2} = 75\%$$

Today’s Agenda

- Testing and Error Metrics
  - Training, Testing
  - Accuracy
  - Precision
  - **Recall**
  - F-Score

Recall

		Diagnosis	
		Diagnosed Sick	Diagnosed Healthy
Patients	Sick	1,000	200
	Healthy	800	8,000

Recall:
Out of all the sick patients, how many did we correctly diagnose as sick?

Recall

Diagnosis

		Diagnosed Sick	Diagnosed Healthy
Patients	Sick	1,000	200
	Healthy	800	8,000

Recall:
Out of all the sick patients, how many did we correctly diagnose as sick?

Recall =

$$\frac{1{,}000}{1{,}000 + 200} = 83.3\%$$

Recall

		Folder	
		Spam Folder	Inbox
Email	Spam	100	170
	Not Spam	30	700

Recall:
Out of all the spam emails, how many were correctly sent to the spam folder?

Recall

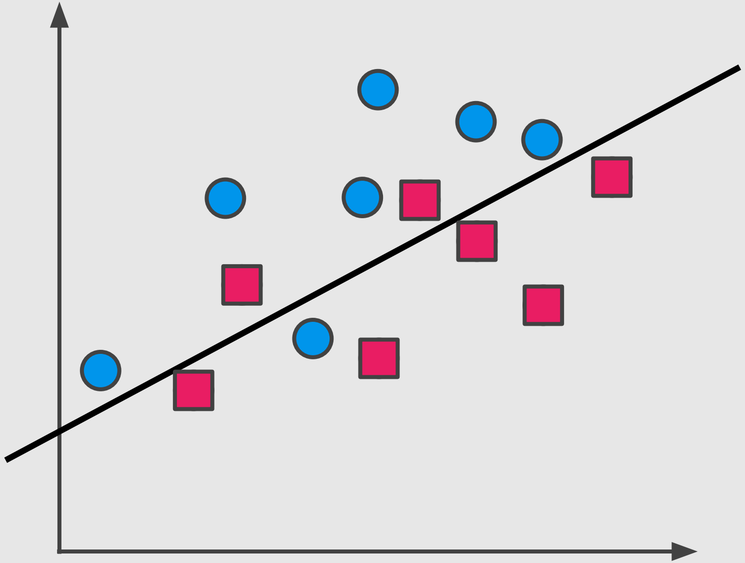
		Folder	
		Spam Folder	Inbox
Email	Spam	100	170
	Not Spam	30	700

Recall:
Out of all the spam emails, how many were correctly sent to the spam folder?

Recall =

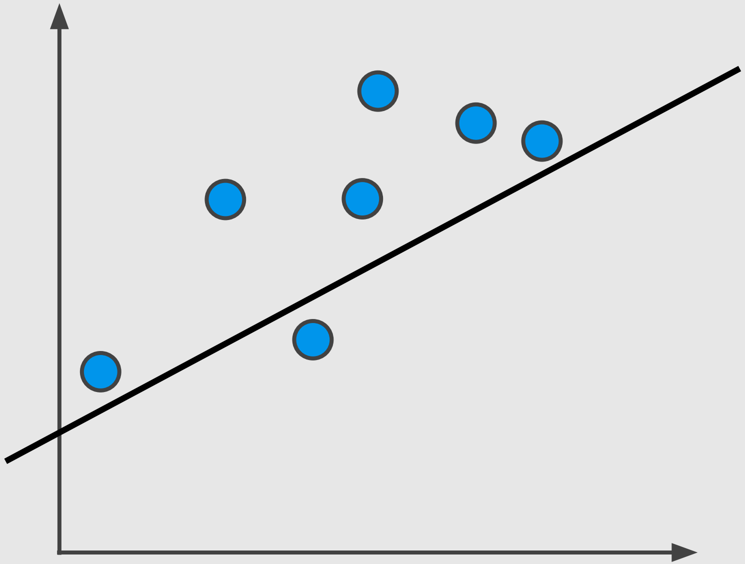
$$\frac{100}{100 + 170} = 37\%$$

Recall



Recall:
Out of all the points labelled positive, how many did we correctly predict?

Recall



Recall:
Out of all the points labelled positive, how many did we correctly predict?

Recall =

$$\frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

$$\frac{6}{6 + 1} = 85.7\%$$

Precision and Recall

Medical Model

Precision: 55.7%

Recall: 83.3%

Spam Detector

Precision: 76.9%

Recall: 37%

One Score?

Medical Model

Precision: 55.7%

Recall: 83.3%

Average = 69.5%

Spam Detector

Precision: 76.9%

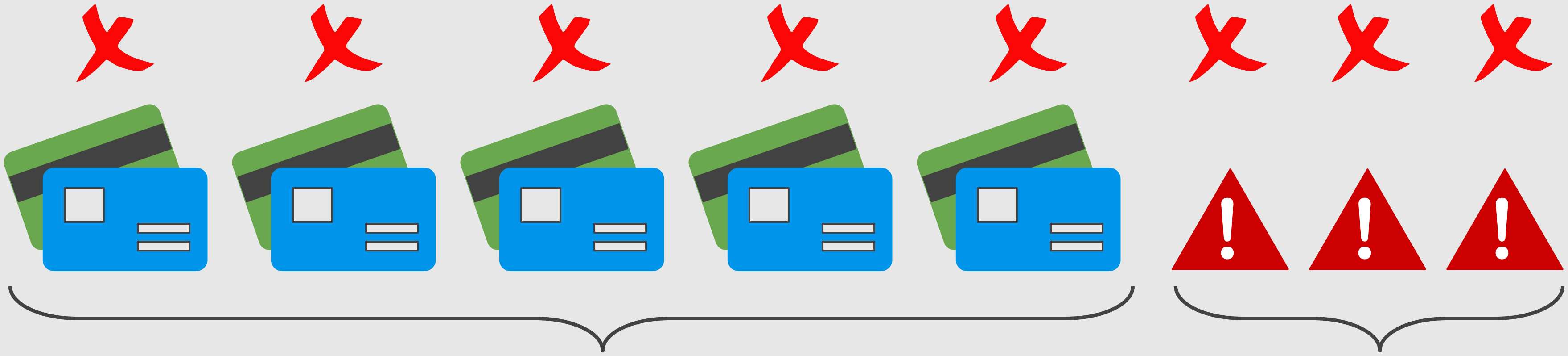
Recall: 37%

Average = 56.9%

Today's Agenda

- Testing and Error Metrics
  - Training, Testing
  - Accuracy
  - Precision
  - Recall
  - **F-Score**

Credit Card Fraud

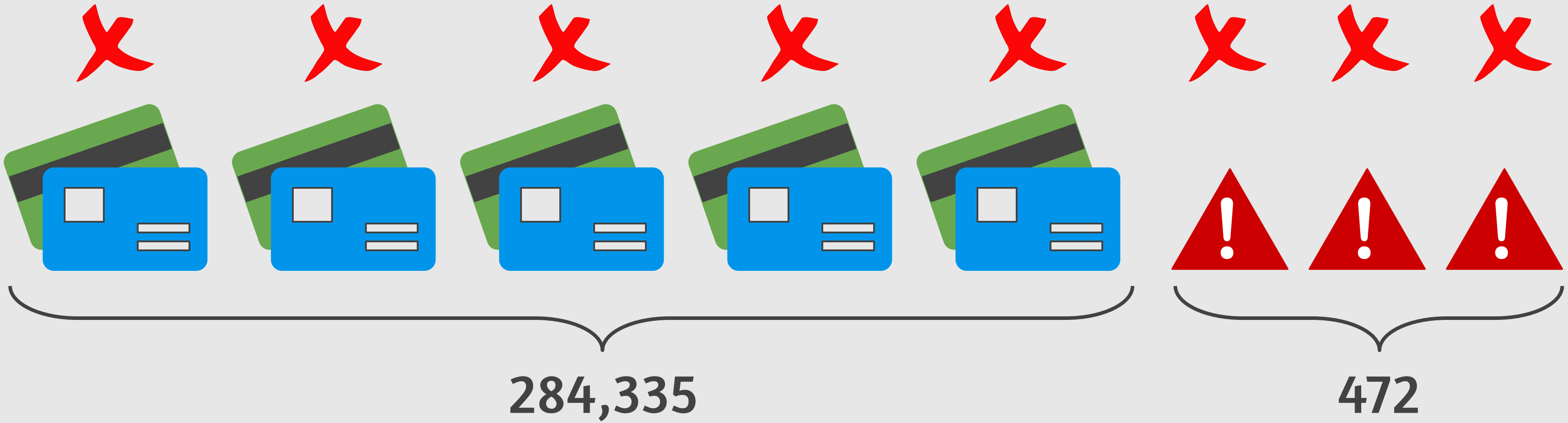


284,335

472

Model: All transactions are fraudulent.

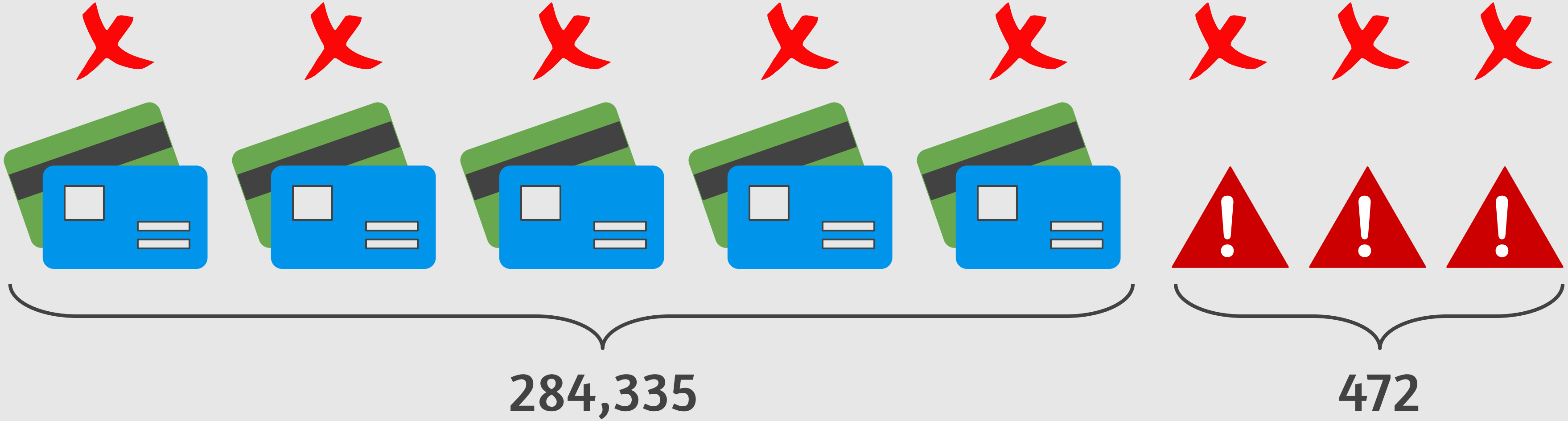
Credit Card Fraud



Model: All transactions are fraudulent.

$$\text{Precision} = \frac{472}{284{,}807} = 0.016\%$$

Credit Card Fraud

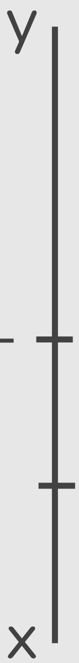


Model: All transactions are fraudulent.

$$\text{Precision} = \frac{472}{284{,}807} = 0.016\% \qquad \text{Recall} = \frac{472}{472} = 100\%$$

Harmonic Mean

$$\text{Arithmetic Mean} = \frac{x + y}{2}$$



Harmonic Mean

$$\text{Arithmetic Mean} = \frac{x + y}{2}$$

$$\text{Harmonic Mean} = \frac{2xy}{x + y}$$

Precision: 1

Recall: 0

Average = 0.5

Harmonic Mean = 0

Precision: 0.2

Recall: 0.8

Average = 0.5

Harmonic Mean = 0.32

F1 Score = Harmonic Mean (Precision, Recall)

F1 Score

Medical Model

Precision: 55.7%

Recall: 83.3%

Average = 69.5%

F1 Score = 66.8%

F1 Score

Spam Detector

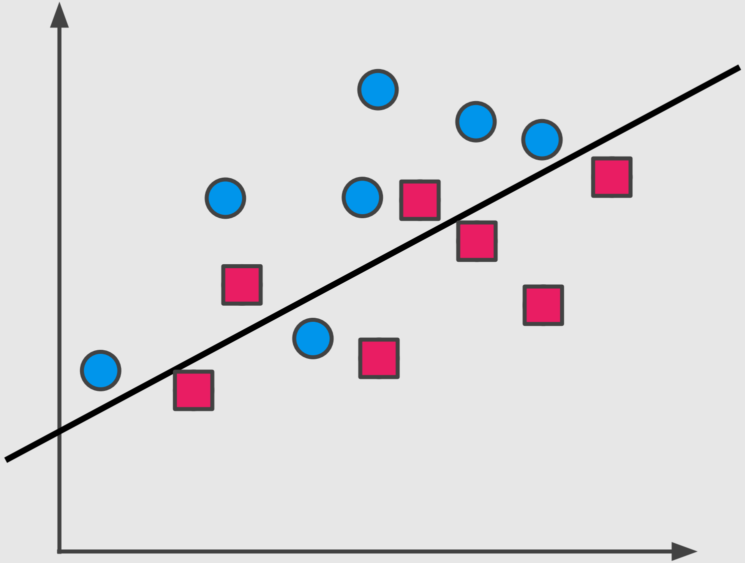
Precision: 76.9%

Recall: 37%

Average = 56.9%

F1 Score = 50.0%

F1 Score



Precision: 75%

Recall: 85.7%

Average = 80.3%

F1 Score = 80%

F_β Score

F_β Score

Precision

Recall

F_β Score

Precision F0.5 Score F1 Score F2 Score Recall

F_β Score

?

Precision F0.5 Score F1 Score F2 Score Recall

F_β Score

F10 Score

Precision F0.5 Score F1 Score F2 Score Recall

F_β Score

F1 Score = Harmonic Mean (Precision, Recall)

Fβ Score

F1 Score = Harmonic Mean (Precision, Recall)

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

Fβ Score

F1 Score = Harmonic Mean (Precision, Recall)

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

$$F_1 = 2 \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

F$_\beta$ Score

$$F_1 = 2\,\frac{\text{precison}\cdot\text{recall}}{\text{precision} + \text{recall}}$$

$$F_\beta = (1 + \beta^2)\,\frac{\text{precison}\cdot\text{recall}}{(\beta^2\cdot\text{precision}) + \text{recall}}$$

Today’s Agenda

- Testing and Error Metrics
  - Training, Testing
  - Accuracy
  - Precision
  - Recall
  - F-Score

References

- https://scikit-learn.org/stable/modules/model_evaluation.html
- https://en.wikipedia.org/wiki/Precision_and_recall
- https://en.wikipedia.org/wiki/Binary_classification
- https://en.wikipedia.org/wiki/F1_score
- https://www.quora.com/What-is-an-intuitive-explanation-of-F-score
- "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms", Neural Computation, 1998

Machine Learning Courses

- "Testing and Error Metrics" https://youtu.be/aDW44NPhNw0
- "ROC Curve" https://youtu.be/z5qA9qZMyw0