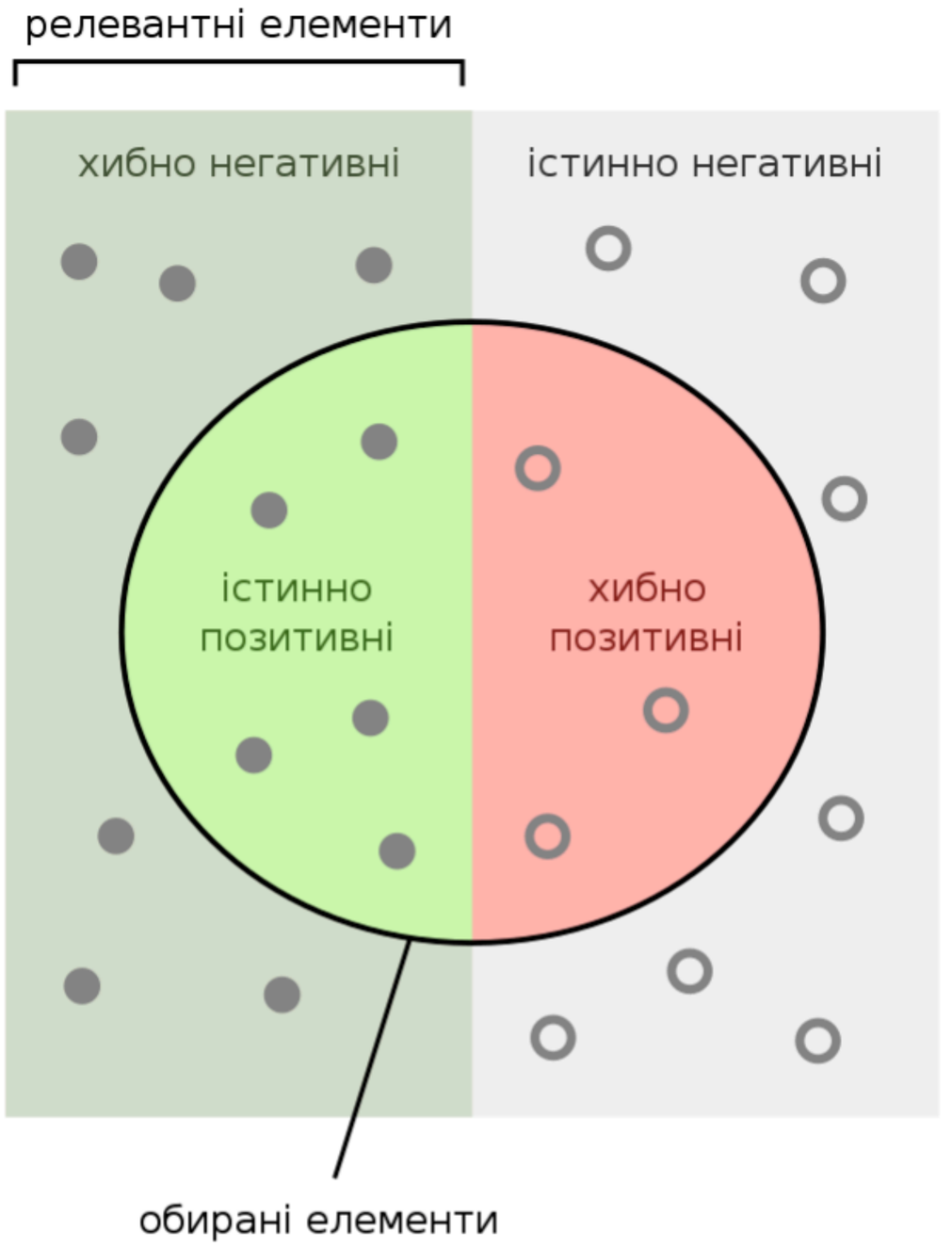
Оцінювання

Оцінювання моделей машинного навчання



Як багато з обираних елементів є релевантними?

$$\text{Влучність} = \frac{\text{істинно позитивні}}{\text{обирані елементи}}$$

Як багато з релевантних елементів стають обраними?

$$\text{Повнота} = \frac{\text{істинно позитивні}}{\text{релевантні елементи}}$$

		Actual	
		Positive	Negative
Predicted	Positive	**True Positive**	**False Positive**
	Negative	**False Negative**	**True Negative**

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

$$\text{accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fp} + \text{fn}}$$

$$\text{F}_1 \text{ score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Accuracy
Точність

Accuracy - яку частину з розміченого корпусу модель вгадала?

Проблема. Коли є дуже багато true negatives і відносно мало true positives, то результати можуть бути оманливо високими.

Приклад з NER. Припустімо, що в золотому корпусі є 180 слів, які не є званою сутністю, і 20 слів, які є званою сутністю. Припустімо, що наша модель не розпізнала жодної званої сутності. В такому випадку у нас є 180 true negatives, 0 true positives, 20 false negatives і 0 false positives. Accuracy буде (180+0)/(180+0+20+0) - 90%!

АЛЕ! І recall, і precision у такому випадку буде 0.

Precision

Теж точність :) (влучність)

Precision - скільки **мусору** повернула модель?

Проблема. Precision не дивиться на пропущені правильні відповіді (false negatives). Тому якщо модель повернула мало мусору але також **замало** відповідей в загальному, precision цього не помітить.

Приклад. З 20 іменованих сутностей модель повернула лише 5, але всі вони були дійсно іменованими сутностями. В такому випадку матимемо true positive 5, false positive 0, true negative 180, false negative 15. Precision буде 100%, а accuracy 92,5%!

АЛЕ recall буде всього 25%.

Recall
Покриття

Recall - інтуїтивно, скільки правильних відповідей модель **пропустила**.

Проблема. Recall не зважає на false positives. Тому якщо їх дуже багато, модель може бути зовсім непродуктивною, але recall буде дуже високим.

Приклад. Припустімо, наша модель позначає всі слова як звані сутності. Тоді маємо true positive 20, false positive 180, true negative 0, false negative 0. Recall буде 100%!

АЛЕ! Accuracy буде 20%, а precision - 20%

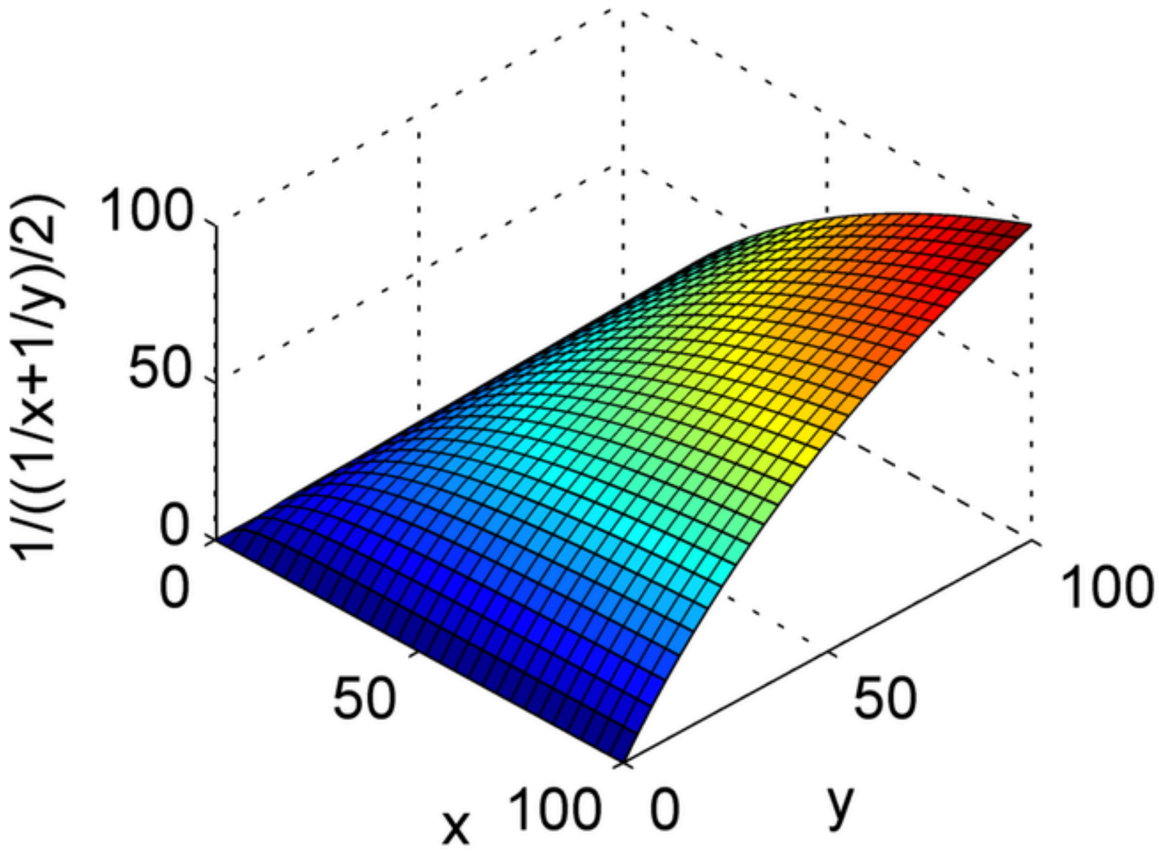
F1 Score

Harmonic mean of precision and recall

$$F_1\text{–score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

F1 - це середнє гармонійне точності і покриття. Інтуітивно, його використовують, щоб відобразити наскілки збалансовані ці два показники.

Справа в тому, що якщо одне зі значень високе, а друге низьке, то середнє *арифметичне* поверне значення посередині. Середнє *гармонійне* поверне високе значення, якщо обидва значення високі, низьке, якщо обидва низькі, **але** якщо одне зі значень значно нижче, ніж інше, то повернеться число, наближене до низького.



Confidence Pipeline

Dictionary-based	very high precision, very low recall
Rule-based	high precision, low recall (doesn't return previously unseen entities)
Classifier trained on the wrong data	meh precision, ok recall
Classifier trained on the right data	ok precision, high recall (returns previously unseen entities)
Classifier tuned on your data	high precision, high recall (but needs data!)