Build a Student Intervention System

* Classifier choices & comparisons from lectures

→ Decision trees

→ Parametric vs Instance Based Learners pg 13

- biased
- known functional relation as an initial guess

- unbiased
- unknown relationship

? Linear Regression KNN ← perf comparison pg 29

→ Neural Networks: Perceptron vs Gradient Descent pg 17

- guaranteed convergence for linearly separable data
- more robust in the case of non-linearly separable data

- local optima
- advanced optimisation methods pg 19

→ Support Vector Machine

→ Naive Bayes

* Discussion points re. performance

METRICS
→ Train vs Query performance KNN vs linear regression
→ F1 score - want to be conservative - intervention is preferable rather than to be avoided
- baselines ?

→ Opt computing - approximate solutions
- quality improves indefinitely with learning

Comparing performance with complexity for individual algorithms e.g.
→ Data size
→ Decision Tree depth query time ?

* "find the most effective model with the least amount of computation costs (you pay the company by the memory and CPU time you use on their servers)"

Model Evaluation:

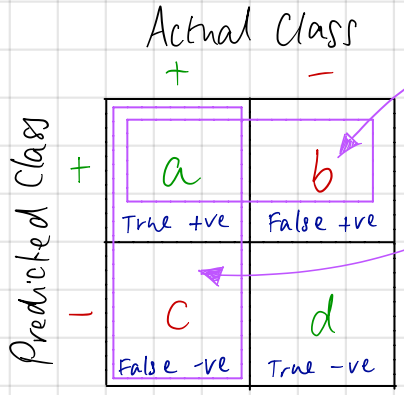
1. F1 score
2. Training set size
3. Prediction computational cost

} which do we do more of?
how often do we get more data to train

* Naive Bayes in Laymens terms:

Bayes Theorem:

* Confusion matrix



precision $= a/(a+b) = \dfrac{\#\text{ true +ve}}{\#\text{ true +ve} + \#\text{ false +ve}} \leq 1$

recall $= a/(a+c) = \dfrac{\#\text{ true +ve}}{\#\text{ true +ve} + \#\text{ false -ve}} \leq 1$

$$F1 = \frac{2\,\text{precision}\,\text{recall}}{\text{precision} + \text{recall}}$$

$$F1 = \frac{2\dfrac{\#\text{ true +ve}}{\#\text{ true +ve} + \#\text{ false +ve}}\dfrac{\#\text{ true +ve}}{\#\text{ true +ve} + \#\text{ false -ve}}}{\dfrac{\#\text{ true +ve}}{\#\text{ true +ve} + \#\text{ false +ve}} + \dfrac{\#\text{ true +ve}}{\#\text{ true +ve} + \#\text{ false -ve}}}$$

$$= \frac{2(\#\text{ true +ve})^2}{\#\text{ true +ve}\,(2\,\#\text{ true +ve} + \#\text{ false +ve} + \#\text{ false -ve})}$$

$$= \frac{2\,\#\text{ true +ve}}{(2\,\#\text{ true +ve} + \#\text{ false +ve} + \#\text{ false -ve})}$$

bad $= 0 \leq F1 \leq 1 =$ good

■ Baselines for performance metrics

→ We are interested in identifying students which might fail and thus require intervention so choosing 'fail' or 'intervene' as our +ve class gives the more meaningful F1 score

→ For a given accuracy score we prefer false +ves to false -ves so in the best case recall = 1 but in general we want

precision < recall

→ Our data is imbalanced i.e. pass rate ≈ 67% so, ≈ 1/3 of the students require intervention

Let pass rate = p

- suppose we always predict pass

accuracy = $p\ (= \frac{2}{3})$
recall = 0
precision = $\frac{1}{3}$
} F1 = 0

0	0
1-p	p

- suppose we always predict fail

accuracy = precision = $1-p\ (= \frac{1}{3})$
recall = 1
$\Rightarrow F1 = \frac{2(1-p)}{2-p} \quad \left(= \frac{2 \cdot \frac{1}{3}}{4/3} = \frac{1}{2}\right)$

1-p	p
0	0

- randomly choose pass/fail — choose pass q of the time

$\Rightarrow$ accuracy $= (1-q)(1-p) + qp$

$= 1 - q - p + 2qp$

$= \frac{1}{3} - q\left(1 - \frac{4}{3}\right) = \frac{1}{3}(1+q)$

Predicted Class \ Actual Class	1	0
1	$(1-q)(1-p)$ True +ve	$(1-q)p$ False +ve
0	$q(1-p)$ False -ve	qp True -ve

$\Rightarrow$ precision $= \frac{(1-q)(1-p)}{(1-q)(1-p+p)} = 1-p$

$\Rightarrow$ recall $= \frac{(1-q)(1-p)}{(1-p)(1-q+q)} = 1-q$

$\Rightarrow F1 = \frac{2(1-p)(1-q)}{2-p-q} = \frac{\frac{2}{3}(1-q)}{\frac{4}{3}-q} = \frac{2(1-q)}{4-3q}$

Now for a given accuracy we prefer false +ve to false -ve. We note $p=q$ gives

precision = recall = $\frac{1}{3}$

accuracy = $5/9 > \frac{1}{2}$

$F1 = \frac{\frac{2}{3}}{4-2} = \frac{1}{3}$

* Relationship between accuracy and f1 score

Predicted Class \ Actual Class	+	−
+	a (True +ve)	b (False +ve)
−	c (False −ve)	d (True −ve)

$$\text{accuracy} = \frac{a+d}{a+b+c+d}$$

$$\text{precision} = \frac{a}{a+b} \qquad \text{recall} = \frac{a}{a+c}$$

$$\text{f1 score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

$$f = \frac{2pq}{p+q} \Rightarrow \frac{df}{dp} = \frac{(p+q)2q - 2pq}{(p+q)^2}$$

$$= \frac{2q^2}{(p+q)^2} > 0$$

- Increase accuracy

Case (i) increase a

⇒ decrease either b (or c)
⇒ increase precision (or recall) ← numerator increases, denominator unchanged
⇒ increase f1 score

Case (ii) increase d

⇒ decrease either b (or c)
⇒ increase precision (or recall) ← numerator unchanged, denominator decreases
⇒ increase f1 score

- Fixed accuracy

Case (i) increase a decrease d

⇒ increase precision and recall ← numerator and denominator increase by the same amount
⇒ increase f1 score

In conclusion an increase in accuracy ⇒ increase in f1 score
but
an increase in f1 score ⇏ increase in accuracy