LUMS

CENTER FOR SPEECH AND LANGUAGE TECHNOLOGIES

Evaluation of Classifiers

Which Mean to Use?

Agha Ali Raza

CS535/EE514 – Machine Learning

		Gold Labels			
		Gold Positive	**Gold Negative**		
Predicted Labels	**Predicted Positive**	True Positives (tp)	False Positives (fp)	$\frac{tp}{tp+fp}$	"Precision" aka "Positive Predictive Value"
	Predicted Negative	False Negatives (fn)	True Negatives (tn)	$\frac{tn}{fn+tn}$	"Negative Predictive Value"
		$\frac{tp}{tp+fn}$ "Recall" aka "Sensitivity" aka "True Positive Rate"	$\frac{tn}{fp+tn}$ "Specificity" aka "True Negative Rate"		
		$\frac{fn}{tp+fn}$ 1 - Sensitivity = "False Negative Rate" aka "False Rejection Rate"	$\frac{fp}{fp+tn}$ 1 - Specificity = "False Positive Rate" aka "False Acceptance Rate"		

$$Accuracy = \frac{tp+tn}{tp+fp+tn+fn}$$

True Positive Rate = "Positive Likelihood Ratio"
False Positive Rate

False Negative Rate = "Negative Likelihood Ratio"
True Negative Rate

Probability = "Odds," often expressed as X:Y
1 - Probability

A worked example

You are shown a set of 21 coins: 10 gold and 11 copper. Your task to accept all gold coins and reject all copper ones

You accept 7 coins as being gold (these are your positives)

- 5 of these are actually gold (these are your true positives, tp)
- 2 of these are copper (these are your false positives, fp)
- You falsely rejected 5 gold ones (false negatives, fn)
- You correctly rejected 9 copper ones (true negatives, tn)

	Actual Gold	**Actual Copper**
Predicted Gold	5	2
Predicted Copper	5	9

- Your **precision** is $\frac{tp}{all\ of\ your\ positives} = \frac{tp}{tp+fp} = \frac{5}{7}$
- Your **recall** is $\frac{tp}{number\ of\ actual\ gold\ coins} = \frac{tp}{tp+fn} = \frac{5}{10}$
- Your **specificity** is $\frac{tn}{number\ of\ copper\ coins} = \frac{tn}{fn+tn} = \frac{9}{11}$
- Your **accuracy** is $\frac{correct\ answers}{all\ attempts} = \frac{tp+tn}{tp+tn+fp+fn} = \frac{5+9}{5+9+2+5} = \frac{14}{21}$

Realistic extremes

You accept only one coin and that is gold

- Your precision is very high (1/1) but recall is very low (1/10)

	Ac Gld	Ac Cop
Pr Gld	1	0
Pr Cop	9	11

You return all 21 coins

- Your recall is very high (10/10) but precision is very low (10/21)

	Ac Gld	Ac Cop
Pr Gld	10	11
Pr Cop	0	0

Only one out of the 21 coins is gold. And you reject everything.

- Your accuracy is very high (20/21 = 0.95) but precision/recall are 0

	Ac Gld	Ac Cop
Pr Gld	0	0
Pr Cop	1	20

So, what do we do now?

- A combined measure?

A combined measure of Precision and Recall

- It is useful to have a single number to describe performance. Should be high when both P and R are high.
- The mean of precision and recall?
- But what kind of a mean should we use?
- Simple Arithmetic Mean is problematic: e.g.
  - P = 0.0, R = 1.0, AM = 0.5
  - P = 0.1, R = 0.9, AM = 0.5
- **Requirements:**
  - We need a weighted mean as we may care more about P or R
  - We need a conservative (deliberately lower) estimate of mean
  - If P and R are far apart we need the mean to tend to the lower value
  - In order to do well, a classifier must do well on both P and R so that it cannot beat the system by being either too selective or too indiscriminate

What is mean (or average)?

- The **central tendency** is a single number that represents the most common value for a list of numbers.
  - It is the value that has the highest probability from the probability distribution that describes all possible values that a random variable may have.
- Many ways to calculate it
  - **Mode:** the most common value in the data distribution
  - **Median:** the middle value if all values in the data sample were ordered
  - **Mean:** the average value – Three common types
- The *mean* is different from the *median* and the *mode* in that it is a measure of the central tendency that is calculated from the data.

What is mean (or average)?

- Different ways to calculate the mean based on the type of data.
- Three common types:
  - Arithmetic mean
  - Geometric mean
  - Harmonic mean

  aka, the Pythagorean means

Types of Means

Arithmetic Mean

$$AM = \frac{a_1 + a_2 + a_3 + \cdots + a_n}{n}$$

For 2 values: $$AM = \frac{a_1 + a_2}{2}$$

Geometric Mean

$$GM = \sqrt[n]{a_1 . a_2 . a_3 \ldots a_n}$$

For 2 values: $$GM = \sqrt[2]{a_1 a_2}$$

Harmonic Mean

$$HM = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} + \cdots + \frac{1}{a_n}}$$

For 2 values: $$HM = \frac{2}{\frac{1}{a_1} + \frac{1}{a_2}} = \frac{2a_1 a_2}{a_1 + a_2}$$

$$\frac{2PR}{P+R}$$

Arithmetic Mean

Ref: http://economistatlarge.com/finance/applied-finance/differences-arithmetic-geometric-harmonic-means

- The most common average, simplest to compute
- The AM of 2, 3 and 4 is 3
- Suitable when:
  - the data is not skewed (no extreme outliers)
  - the individual data points are not dependent on each other
  - the numbers are relatively evenly distributed, e.g.
    - follow a normal distribution
    - when you are rolling a fair die, the expected value is the mean of all the numbers on it
- Easily distorted if the sample of observations contains outliers or for data that has a non-Gaussian distribution (e.g. multiple peaks – a multi-modal probability distribution).
- The AM is more meaningful when a variable has a Gaussian or Gaussian-like data distribution.

Geometric Mean

Ref: https://towardsdatascience.com/on-average-youre-using-the-wrong-average-geometric-harmonic-means-in-data-analysis-2a703e21ea0

- The geometric mean multiplies rather than sums values, then takes the *nth* root rather than dividing by *n*
- It essentially says: if every number in our dataset was the same number, what would that number have to be in order to have the same *multiplicative product* as our actual dataset?
- This makes it well-suited for describing multiplicative relationships, such as rates and ratios, even if those ratios are on different scales (i.e. do not have the same denominator)
- The geometric mean works well when the data is in an multiplicative relationship or in cases where the data is compounded

Geometric Mean

Ref: https://towardsdatascience.com/on-average-youre-using-the-wrong-average-geometric-harmonic-means-in-data-analysis-2a703e21ea0

- Geometric mean should be used when the data points are inter-related
- It's appropriate for numbers that are distributed along a logarithmic scale - that is, when you're as likely to find a number twice the size as a number half the size.
- The geometric mean does not accept negative or zero values, e.g. all values must be positive.
- A fancy feature of the geometric mean is that you can **often** average across numbers on *completely different scales*.

Geometric Mean

- Compare ratings for two coffeeshops using two different sources.
- The problem is that *source* 1 uses a *5-star* scale and *source* 2 uses a *100-point* scale:

Coffeeshop A

- Source 1 rating: 4.5
- Source 2 rating: 68

Coffeeshop B

- Source 1 rating: 3
- Source 2 rating: 75

If we naively take the arithmetic mean of raw ratings for each coffeeshop:

Coffeeshop A = (4.5 + 68) ÷ 2 = 36.25

Coffeeshop B = (3 + 75) ÷ 2 = 39

We'd conclude that **Coffeeshop B** was the winner.

Geometric Mean

- **The right way to do this is:**
- We need to normalize our values onto the same scale before averaging them with the arithmetic mean, to get an accurate result.
- So we multiply the source 1 ratings by 20 to bring them from a 5-star scale to the 100-point scale of source 2:

Coffeeshop A

- 4.5 * 20 = 90
- (90 + 68) ÷ 2 = 79

Coffeeshop B

- 3 * 20 = 60
- (60 + 75) ÷ 2 = 67.5

So we find that **Coffeeshop A** is the true winner, contrary to the naive application of arithmetic mean above.

Geometric Mean

- **For this particular problem**, the geometric mean, allows us to reach the same conclusion:

  Coffeeshop A = square root of (4.5 * 68) = 17.5

  Coffeeshop B = square root of (3 * 75) = 15

- The arithmetic mean is dominated by numbers on the larger scale, which makes us think Coffeeshop B is the higher rated shop. This is because the arithmetic mean expects an additive relationship between numbers and does not account for scales and proportions. Hence the need to bring numbers onto the same scale before applying the arithmetic mean.
- The geometric mean, on the other hand, can handle varying proportions with ease, due to it's multiplicative nature. This is a tremendously useful property, but **we no longer have any interpretable scale at all**. The geometric mean is effectively unitless in such situations.

Harmonic Mean

Ref: http://economistatlarge.com/finance/applied-finance/differences-arithmetic-geometric-harmonic-means

- The harmonic mean of a set of numbers is the reciprocal of the arithmetic mean of reciprocals
- Best used in situations where extreme outliers exist in the population

x0	x1	x2	x3	x4	x5	x6	x7	x8	AM	GM	HM
1	2	3	4	5	6	7	8	9	5.00	4.15	3.18
2	4	8	16	32	64	128	256	512	113.56	32.00	9.02
5	5	5	5	5	5	5	5	5	5.00	5.00	5.00
5	5	5	5	5	5	5	5	**10**	5.56	5.40	5.29
5	5	5	5	5	5	5	5	**100**	15.56	6.97	5.59
5	5	5	5	5	5	5	5	**1000**	115.56	9.01	5.62
5	5	5	5	5	5	5	5	**10000**	1115.56	11.63	5.62
5	5	5	5	5	5	5	5	**100000**	11115.56	15.03	5.62
5	5	5	5	5	5	5	100000	**100000**	22226.11	45.16	6.43
5	5	5	5	5	100000	100000	100000	**100000**	44447.22	407.89	9.00
5	100000	100000	100000	100000	100000	100000	100000	**100000**	88889.44	33274.21	44.98
100000	100000	100000	100000	100000	100000	100000	100000	**100000**	100000.0	100000.0	100000.0

A combined measure for Prec Rec

Use harmonic mean instead of arithmetic mean as we are taking the average of ratios

Harmonic mean *punishes* extreme values more strictly:

- Consider a *trivial* classifier (always returns class A)
- A very large number of data elements of class B, and a single element of class A:
  - Precision: 0.0
  - Recall: 1.0
  - Arithmetic mean = 0.5 (50% correct), despite being the *worst* possible outcome!
  - The harmonic mean is nearly 0.
  - To have a high F-score, you need *both* a high precision and high recall

F-1-MEASURE

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

$$F = \frac{2}{\frac{1}{P} + \frac{1}{R}} = \frac{2PR}{P + R}$$

F-β-MEASURE

We can choose to favor precision or recall by using an interpolation weight α:

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

$$= \frac{(\beta^2+1)PR}{\beta^2 P + R} \qquad \beta = \frac{1-\alpha}{\alpha}$$

- Balanced F1 measure has $\beta = 1$ (that is, $\alpha = ½$) as shown above.
- To give more weight to the Precision, we pick a β value in the interval
  - $0 < \beta < 1$.
  - Notice that it is getting multiplied with P in the denominator
- To give more weight to the Recall, we pick a β Value in the interval
  - $1 < \beta < +\infty$
- β -> 0 considers only precision, β -> +∞ only recall

F-MEASURE

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

The β parameter differentially weights the importance of recall and precision

- based on the needs of an application
- values of $\beta > 1$ favor recall, while
- values of $\beta < 1$ favor precision

When $\beta = 1$, precision and recall are equally balanced

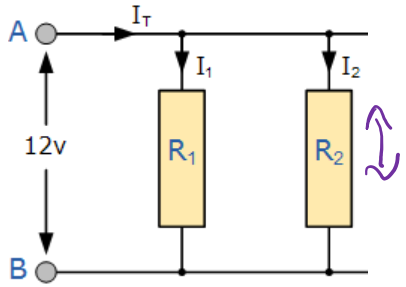
This is the most frequently used metric, and is called F_β=1 or just $F1$:

$$F_1 = \frac{2PR}{P + R}$$

Intuition

Ref: https://stats.stackexchange.com/questions/49226/how-to-interpret-f-measure-values, http://www.electronics-tutorials.ws/resistor/res_4.html

The formula for F-measure (F1, with β=1) is the same as the formula giving the equivalent resistance composed of two resistances placed in parallel in physics (forgetting about the factor 2).



$$\frac{1}{R_T} = \frac{1}{R_1} + \frac{1}{R_2}$$

$$R_T = \frac{R_1 \times R_2}{R_1 + R_2}$$

$$F1 = \frac{2PR}{P+R}$$

This analogy would define F-measure as the equivalent resistance formed by *sensitivity* and *precision* placed in parallel.

Net resistance gets reduced as soon as any one among the two looses resistance

For more details please visit

http://aghaaliraza.com

Thank you!