Performance Metrics for Machine Learning

Sargur N. Srihari
srihari@cedar.buffalo.edu

Topics

1. Performance Metrics
2. Default Baseline Models
3. Determining whether to gather more data
4. Selecting hyperparamaters
5. Debugging strategies
6. Example: multi-digit number recognition

Topics in Performance Metrics

1. Metrics for Regression: squared error, RMS
2. Metric for Density Estimation: KL divergence
3. Metrics for Classification: Accuracy
4. Metrics for Unbalanced data:
   – Loss, Specificity/Sensitivity
5. Metrics for Retrieval: Precision and Recall
6. Combining Precision and Recall: F-Measure
7. Metrics for Image Segmentation: Dice Coefficient

Metrics for Regression

- Sum of squares of the errors between the predictions $y(\boldsymbol{x}_n, \boldsymbol{w})$ for each test data point $\boldsymbol{x}_n$ and target value t_n

$$\boxed{E(\boldsymbol{w}) = \sum_{n=1}^{N} \{y(\boldsymbol{x}_n, \boldsymbol{w}) - t_n\}^2}$$

  - where $\boldsymbol{w}$ are M parameter weights such as those associated with M basis functions

- RMS error

$$\boxed{E_{RMS} = \sqrt{2E(\boldsymbol{w}) / N}}$$

Metric for Density estimation

- K-L Divergence
  - information required as a result of using $q(\mathrm{x})$ in place of $p(\mathrm{x})$

$$KL(p\,||\,q) = -\int p(x)\ln q(x)\,dx - \left(\int p(x)\ln p(x)\,dx\right)$$
$$= -\int p(x)\ln\left\{\frac{p(x)}{q(x)}\right\}dx$$

- Not a symmetrical quantity: $KL(p||q) \neq KL(q||p)$
- K-L divergence satisfies $KL(p||q)>0$ with equality iff $p(x)=q(x)$

Metric for Classification

- For classification and transcription we often measure accuracy of the model
- Accuracy is proportion of examples for which the model produces the correct output
- Error rate: proportion of examples for which model produces an incorrect output
- Error rate is referred to as expected 0-1 loss
  - 0 if correctly classified and 1 if it is not

Loss Function

- Sometimes it is more costly to make one kind of mistake than another
- Ex: email spam detection
  - – Incorrectly classifying legitimate message as spam
  - – Incorrectly allowing a spam message to appear in in box
- Assign higher cost to one type of error
  - – Cost of blocking legitimate message is higher than allowing spam messages

Summary of Loss Functions

- Given a prediction (p) and a label (y), a loss function measures the discrepancy between the algorithm's prediction and the desired output. Squared loss is the default

Loss	Function	Minimizer	Example usage
Squared	$\frac{1}{2}(p-y)^2$	Expectation (mean)	Regression *Expected return on stock*
Quantile	$\tau(y-p)\mathbb{I}(y \geq p)+(1-\tau)(p-y)\mathbb{I}(y \leq p)$	Median	Regression *What is a typical price for a house?*
Logistic	$\log(1+\exp(-yp))$	Probability	Classification *Probability of click on ad*
Hinge	$\max(0, 1-yp)$	0-1 approximation	Classification *Is the digit a 7?*
Poisson		Counts (Log Mean)	Regression *Number of call events to call center*
Classic	Squared loss without importance weight aware updates	Expectation (mean)	Regression *squared loss often performs better than classic.*

https://github.com/JohnLangford/vowpal_wabbit/wiki/Loss-functions

Precision and Recall

- Definitions for binary classification

	Correct Label=T	Correct Label=F
Classifier Label=T	TP	FP Type1 error
Classifier Label=F	FN Type 2 error	TN

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

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Recall} = \frac{TP}{TP+FN}$$

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

- Classification examples

Sample #	Correct Label	Classifier 1 Label
1	F	F
2	F	F
3	F	F
4	F	F
5	F	T
6	T	T

	Correct Label=T	Correct Label=F
Classifier Label=T	1 (TP)	1 (FP)
Classifier Label=F	0 (FN)	4 (TN)

$$\text{Accuracy} = 5/6 = 83\%$$
$$\text{Precision} = 1/2 = 50\%$$
$$\text{Recall} = 1/1 = 100\%$$
$$\text{F-measure} = 2/3 = 66\%$$

Classifier 2 is dumb: always outputs F.
Yet has same accuracy as Classifier 1

Sample #	Correct Label	Classifier 2 Label
1	F	F
2	F	F
3	F	F
4	F	F
5	F	F
6	T	F

	Correct Label=T	Correct Label=F
Classifier Label=T	0 (TP)	0 (FP)
Classifier Label=F	1 (FN)	5 (TN)

$$\text{Accuracy} = 83\%$$
$$\text{Precision} = 0/0 = ?$$
$$\text{Recall} = 0/1 = 0\%$$
$$\text{F-measure} = ?$$

Precision and Recall are useful when the true class is rare, e.g., rare disease.
Same holds true in information retrieval when only a few of a large no. of documents are relevant

Precision-Recall in IR

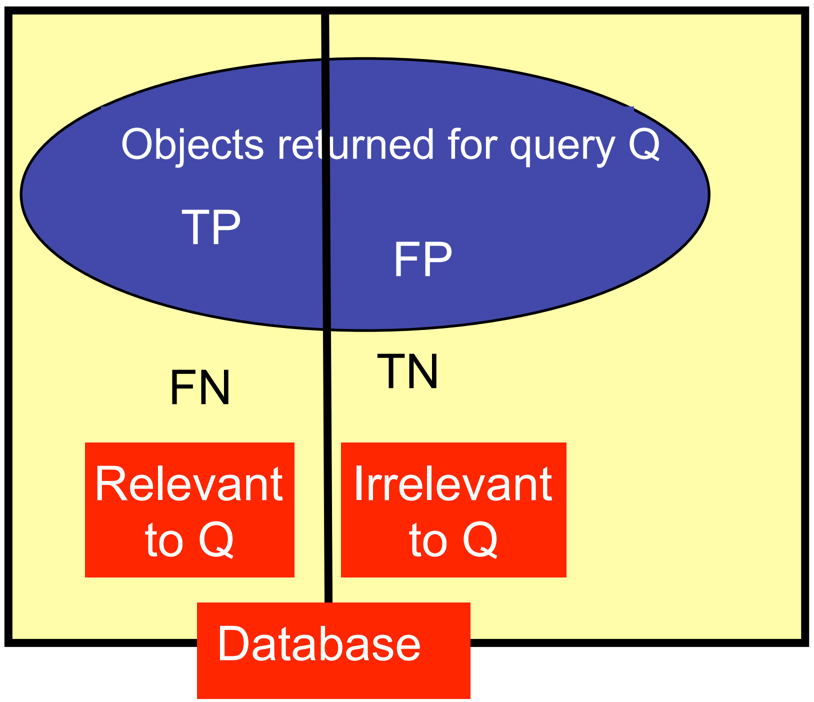
Precision-Recall are evaluated w.r.t. a set of queries

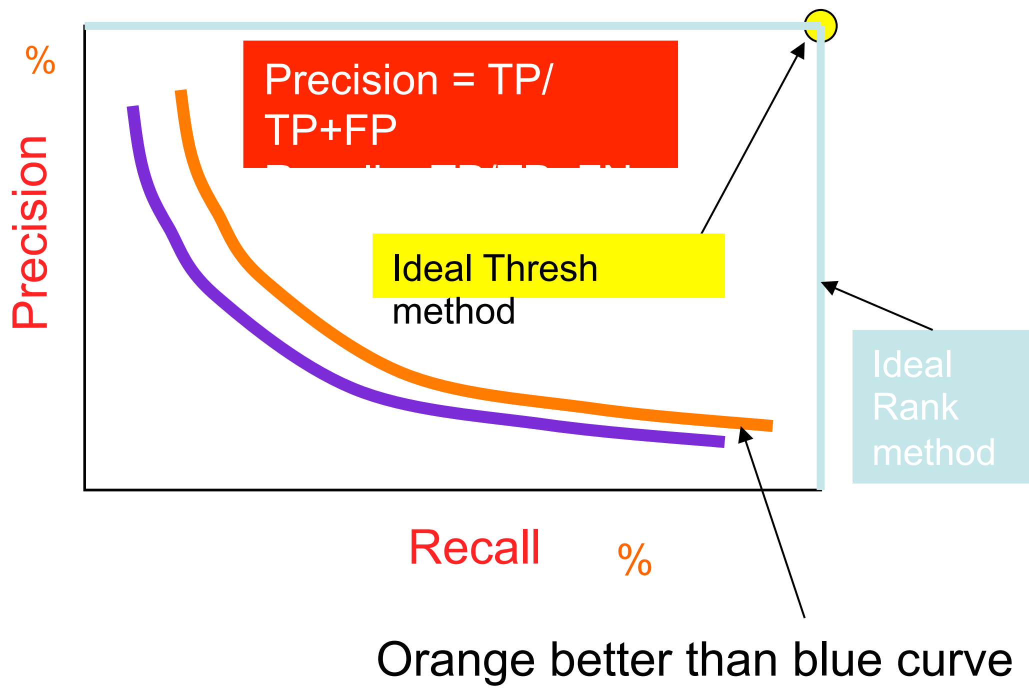
Precision-Recall Curve

Thresh method: threshold *t* on similarity measure

Rank Method: no of top choices presented

Typical inverse relationship

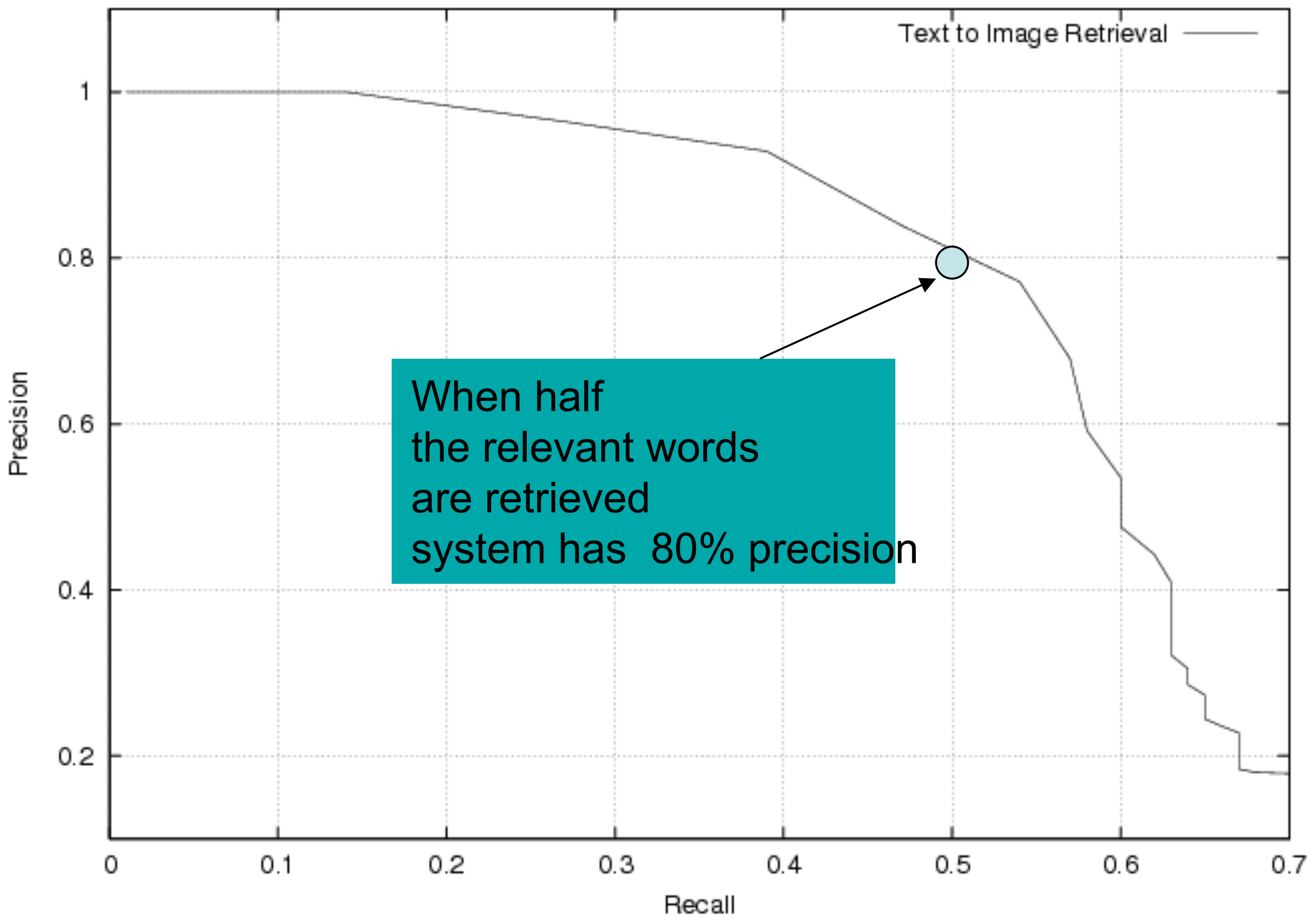




Text to Image search

Experimental settings:

- 150 x 100 = 15,000 word images
- 10 different queries
- Each query has 100 relevant word images



Combined Measures of Precision-Recall

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

Harmonic mean of precision and recall
High value when both P and R are high

$$E = 1 - \frac{1}{\frac{u}{P} + \frac{1-u}{R}} = 1 - \frac{PR}{(1-u)P + uR}$$

u = measure of relative importance of P and R

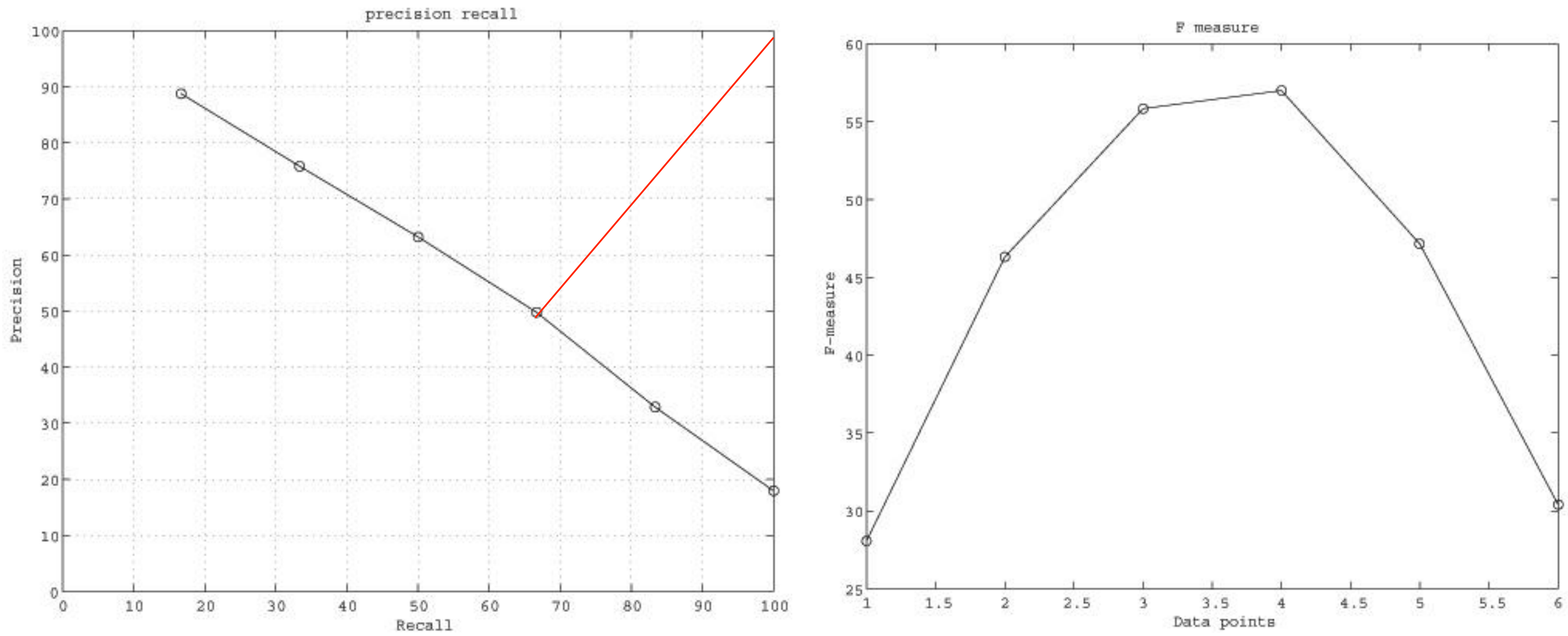
$$u = 1/(v^2+1)$$

The coefficient u has range [0,1] and can be equivalently written as $E = 1 - \frac{(v^2+1)PR}{v^2P+R}$

E-measure reduces to F-measure when precision and recall are equally weighted, i.e. v=1 or u=0.5

$$F = 1 - E = \frac{(v^2+1)PR}{v^2P+R} = \frac{2PR}{P+R}$$

Example of Precision/Recall curve and F-measu



Best F-measure value is obtained when recall = 67% and precision = 50%

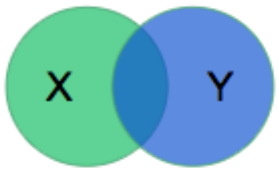
Arabic word spotting

Metric for Image Segmentation

- Dice Coefficient

  X = ROI output by model, a mask

  Y = ROI produced by human expert



$$\mathrm{dice}(X, Y) = \frac{2X \cap Y}{X + Y}$$

Metric is (twice) the ratio of the intersection over the sum of areas.
It is 0 for disjoint areas, and 1 for perfect agreement.
E.g., model performance is written as $0.82\ (0.23)$,
where the parentheses contain the standard deviation.