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

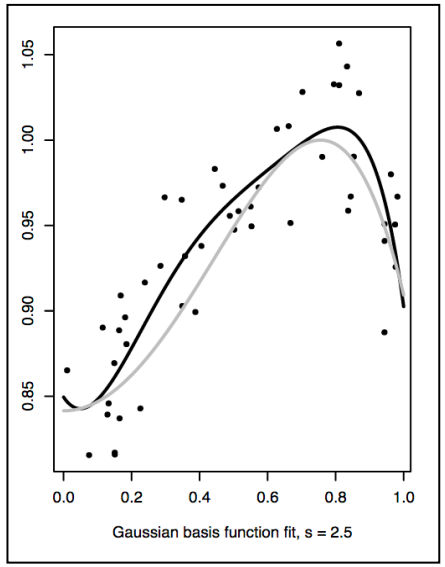
Performance Metrics for ML Tasks

1. Regression: Squared error, RMS
2. Classification: Accuracy
   - Unbalanced data: Loss, Specificity/Sensitivity
3. Density Estimation: KL divergence
4. Information Retrieval: Precision-Recall, F-Measure
5. Image Analysis and Synthesis
   1. Image Segmentation: IOU, Dice
   2. Generative Models: Inception Score, Frechet Inception Distance
6. Natural Language Processing
   - Recognizing Textual Entailment
   - Machine Translation: METEOR

Metrics for Regression

- Linear Regression with feature functions

$$y(\mathrm{x},\mathrm{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathrm{x})$$



- Sum of squares between predictions $y(\boldsymbol{x}_n, \boldsymbol{w})$ and targets in $D = \{(\boldsymbol{x}_n, t_n)\}$, $n = 1,..,N$

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

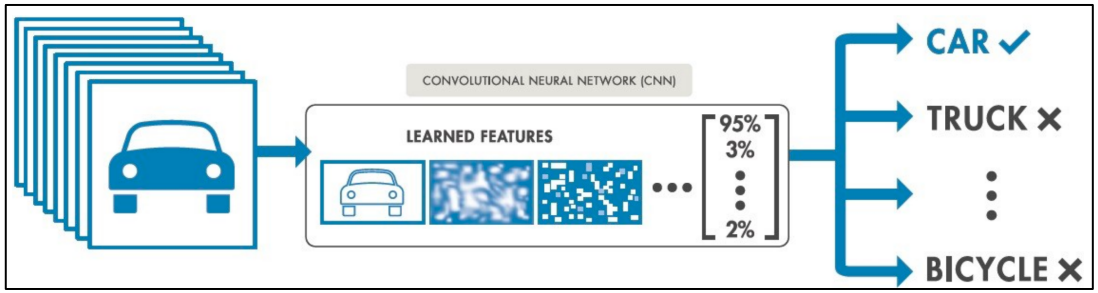
  - where $\boldsymbol{w}$ has M parameters

- RMS error

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

  - Allows comparing different size datasets

Metrics for Classification



- Performance of model measured by

1. Accuracy
   - Proportion of examples for which model produces correct output
2. Error rate
   - Proportion of examples for which model produces incorrect output
     - Error rate is referred to as expected 0-1 loss
     - 0 if correctly classified and 1 if it is not

Loss Function for Classification

- When one kind of mistake costlier than another
  - Ex: email spam detection
    - Incorrectly classifying legitimate message as spam
    - Incorrectly allow spam message into in box
- Assign higher cost to one type of error
  - Ex: Cost of blocking legitimate message is higher than allowing spam messages

Loss for Regression/Classification

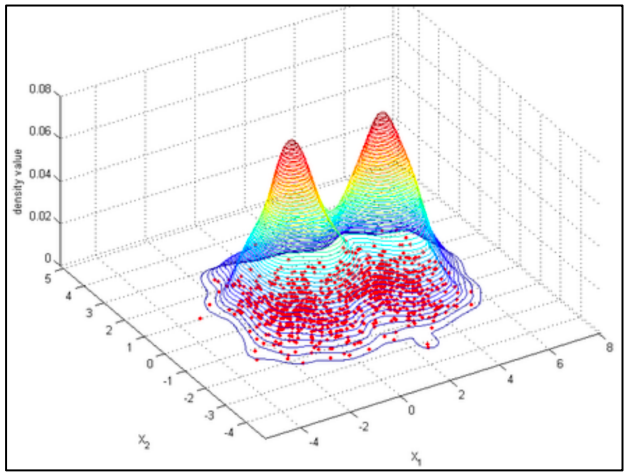
- Given prediction (p) and label (y), a loss function measures the discrepancy between the algorithm's prediction and the desired output.
  - Squared loss is default for regression. Performance metric not necessarily same as Loss.

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

Metric for Density estimation

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



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

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

- Not a symmetrical quantity:

  $KL(p||q) \neq KL(q||p)$

- K-L divergence satisfies $KL(p||q)>0$ with equality iff $p(x)=q(x)$

Information Retrieval: 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

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

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

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

- Compare 2 classifier outputs

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

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)

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

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

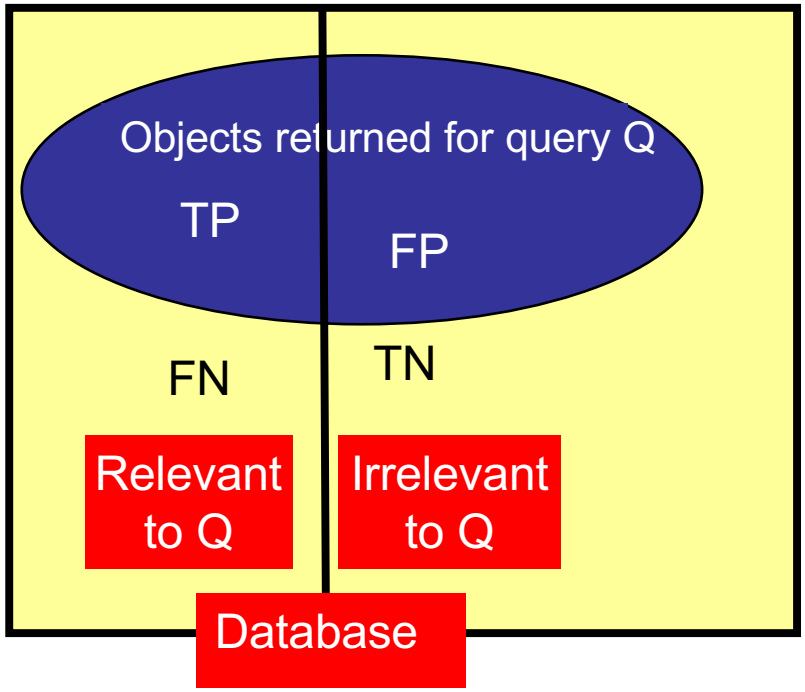
	Correct Label=T	Correct Label=F
Classifier Label=T	0 (TP)	0 (FP)
Classifier Label=F	1 (FN)	5 (TN)

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

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

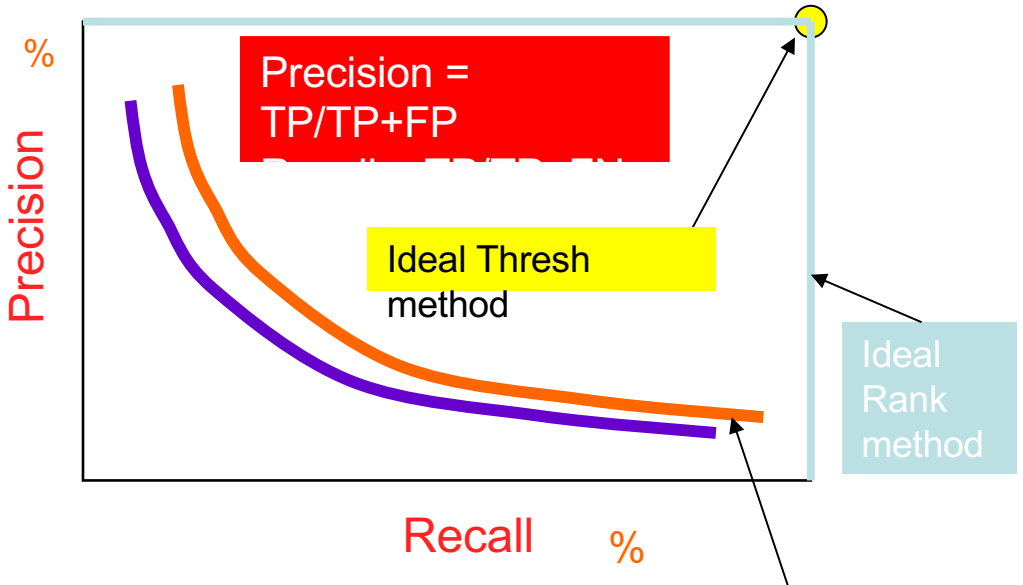


Precision-Recall Curve

Thresh method: threshold *t* on similarity measure

Rank Method: no of top choices presented

Typical inverse relationship

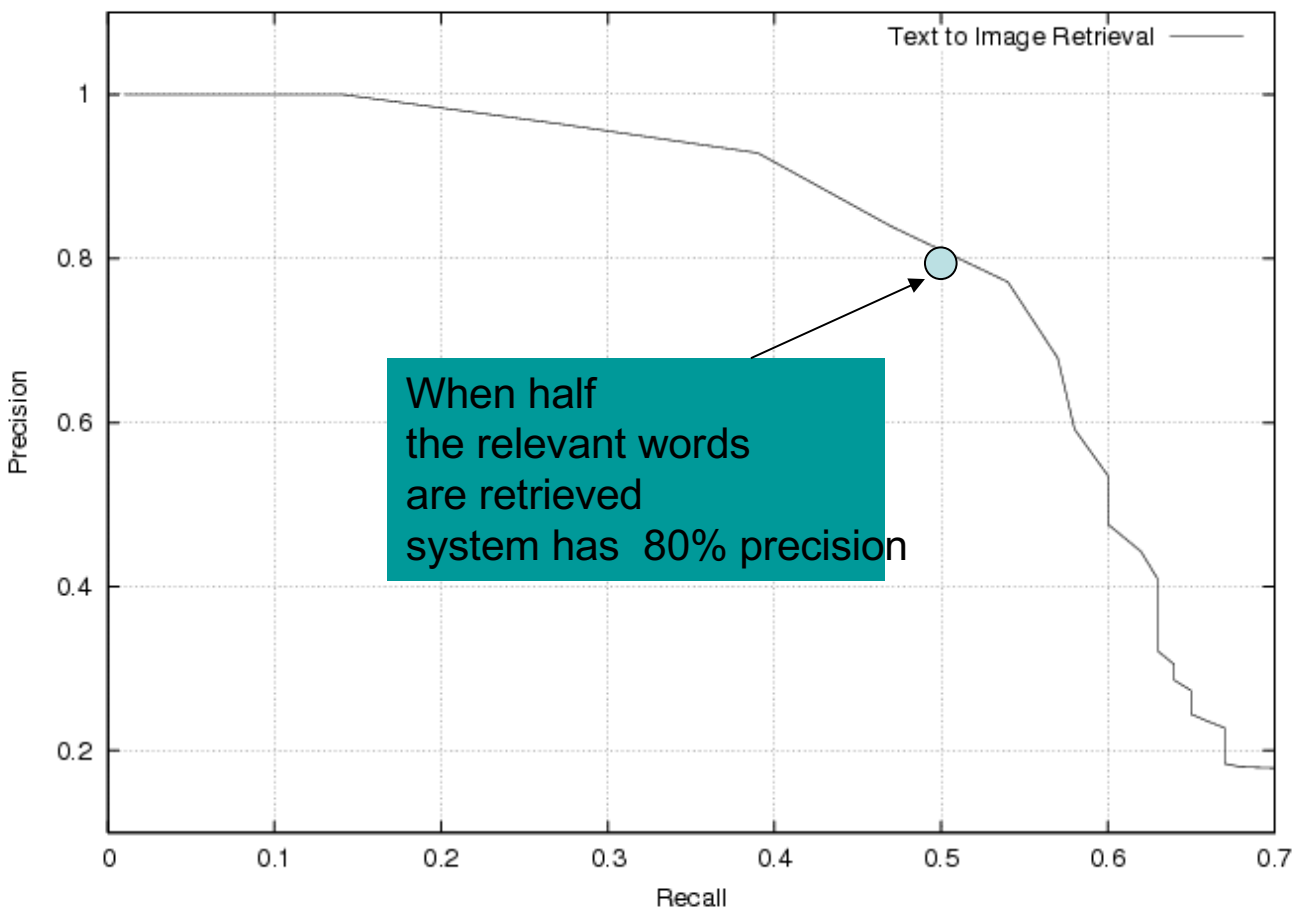


Orange better than blue curve

Text to Image search

Experimental settings:

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



Combined 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

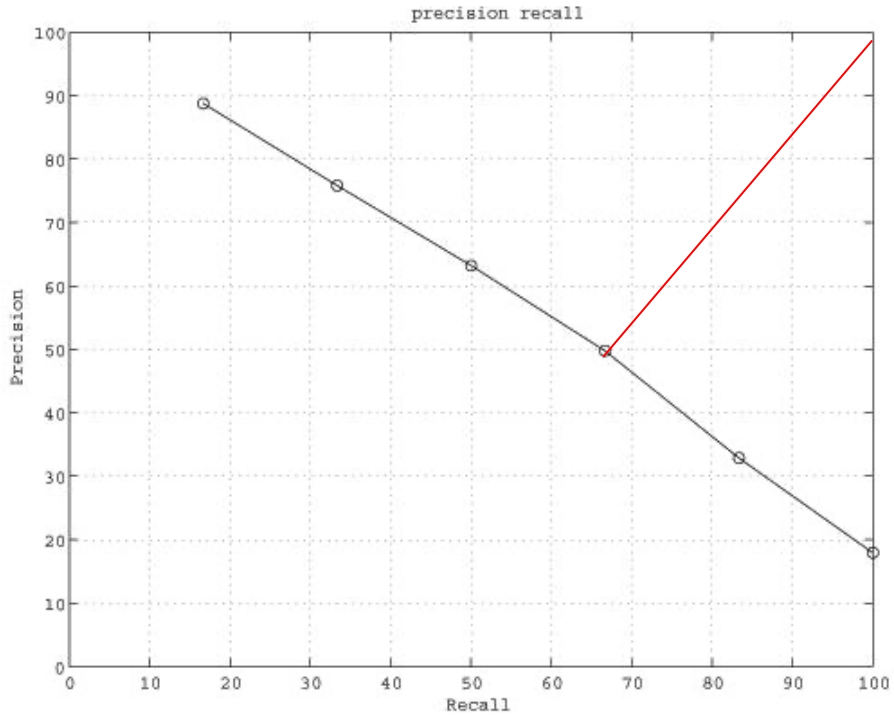
$$u = 1/(v^2+1)$$

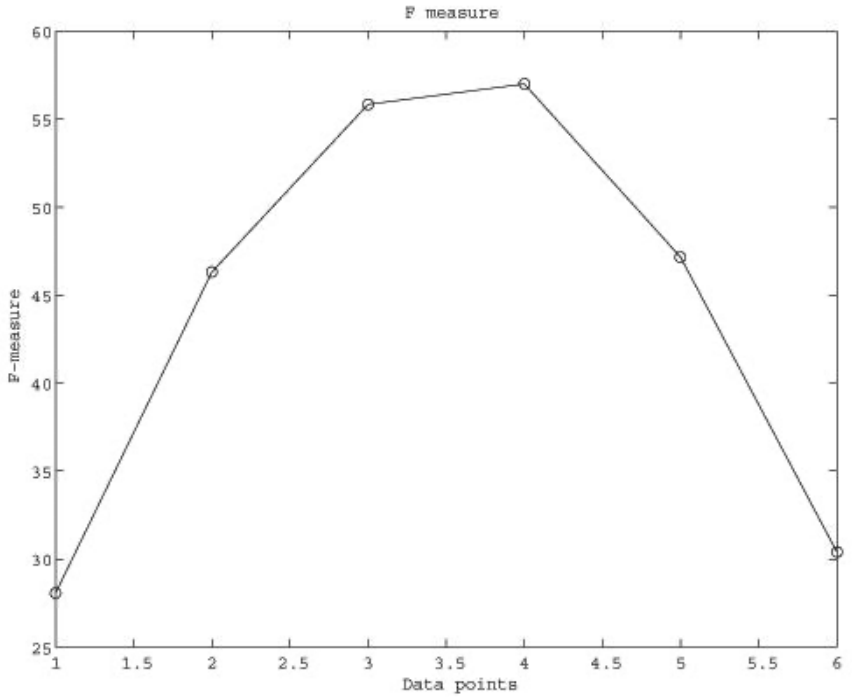
The coefficient u has range [0,1] and can be equivalently written as $E = 1 - \frac{(v^2+1)PR}{v^2P + R}$

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 and F-measure





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

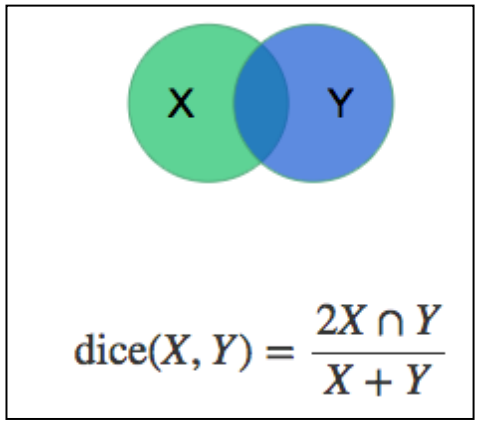
Arabic word spotting

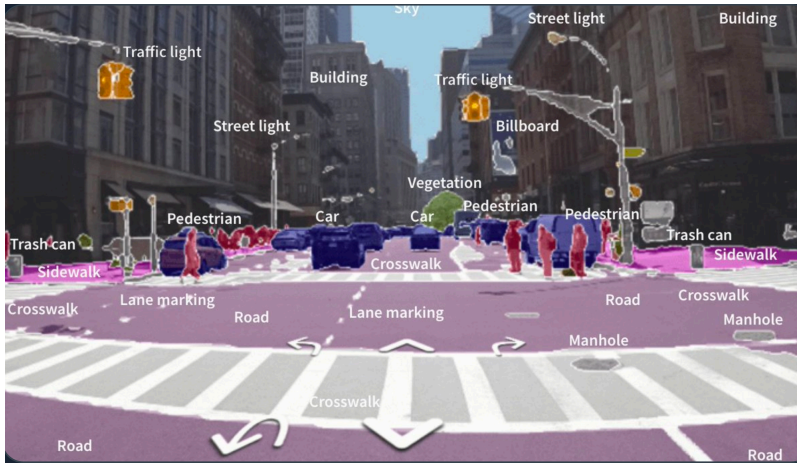
Metric for Image Segmentation

- Dice Coefficient

  X = ROI output by model, a mask

  Y = ROI produced by human expert





Metric is (twice) the ratio of intersection over 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.

Generative Models

The Inception Score (IS) is an objective metric for evaluating the quality of generated images

For synthetic images output by generative adversarial networks

Metrics for Generative Models

- Inception Score (IS) — Intuition
  - InceptionV3 pretrained on ImageNet is used as a robust classifier
  - Inception Score considers two major factors:
    - Diversity and Saliency
      - Diversity is the entropy of the predicted classes between samples, higher diversity (via higher entropy) implies that the generator can produce a broader set of images
        - e.g. if producing images of dogs, it could produce images of many different breeds
      - Saliency is the entropy of the predicted classes within a sample, higher saliency (via lower entropy) implies that the generator is able to produce specific samples belonging to implicit classes
        - e.g. if producing images of dogs, it would generate images of specific breeds rather than blend the features of multiple breeds

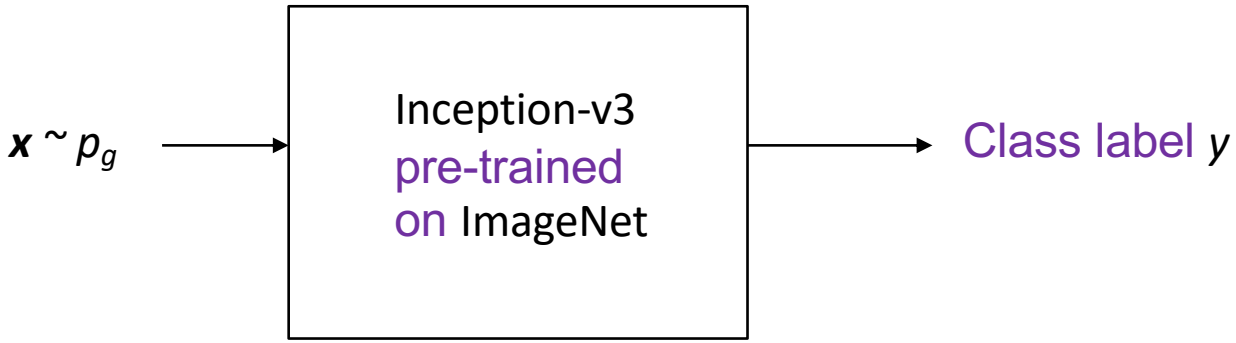
Inception Score (IS) — Formula

IS was the original method for measuring the quality of generated samples. By applying an Inception-v3 network pre-trained on ImageNet to generated samples and then comparing the conditional label distribution with the marginal label distribution:

$$\mathrm{IS}(g) = \exp\left(\mathbb{E}_{x\sim p_g} D_{KL}(p(y|x)\|p(y))\right)$$

- where $x \sim p_g$ indicates an image x is sampled from the generator,
  $p(y|x)$ is the conditional class distribution, and
  $p(y)$ is the marginal class distribution

- IS has a minimum value of zero and a maximum value of infinity, where higher values correspond to better performance



Fréchet Inception Distance (FID)

- Developed as an alternative to Inception Score, the traditional method for measuring the quality of generated images
- Like IS, FID uses an InceptionV3 model pretrained on ImageNet, but they sample from different layers of the network
- IS is a metric which only considers the properties of generated images, whereas FID considers the difference between generated and real images
- In practice, FID is more resistant to noise and is sensitive to mode collapse (artificially pruning modes produces significantly worse results)

Fréchet Inception Distance (FID) — Intuition

- InceptionV3 pretrained on ImageNet is already a very robust classifier, which by extension makes it a very robust feature extractor
- Comparing the extracted features between generated images and real images gives a better underlying idea of the differences which could not be obtained simply by comparing the images directly, or by just examining the generated images
- Use the 2048-dimensional activations of the final pooling layer in a pretrained InceptionV3 network and compare the mean and covariance statistics between generated and real images

Fréchet Inception Distance (FID) — Formula

$$\mathrm{FID}(x, g) = \|\mu_x - \mu_g\|_2^2 + \mathrm{tr}\left(\Sigma_x + \Sigma_g - 2(\Sigma_x \Sigma_g)^{\frac{1}{2}}\right)$$

- Where μ_x and μ_g are the means of the feature maps produced by real and generated images respectively, and Σ_x and Σ_g are the covariance matrices of the feature maps produced by the real and generated images respectively

- FID has a minimum value of zero for when the mean and

Recognizing Textual Entailment

Positive TE:
Text: *If you help the needy, God will reward you*.
Hypothesis: *Giving money to a poor man has good consequences*.

Negative TE:
Text: *If you help the needy, God will reward you*.
Hypothesis: *Giving money to a poor man has no consequences*

Non-TE:
Text: *If you help the needy, God will reward you*.
Hypothesis: *Giving money to a poor man will make you a better person*.

RTE-1 to RTE-5:
•Question answering (QA)
•Relation extraction
•Information retrieval
•Multi-document summarization
•RTE-6 and RTE-7:
Aims at a more natural distribution of positive and negative cases.
•Multi-document summarization
•Update summarization