ARTIFICIAL INTELLIGENCE SOFTWARE DEVELOPMENT

Week 11 Lecture 1

Dr. Hari M Koduvely

Agenda for Today

- ❑ Theory:
    - Generation of Training Data
    - Sampling Techniques
    - Class Imbalance Problems
    - Data Augmentation Techniques

Generation of Training Data

- Production data is messy, noisy and could be unbalanced
- If not handled properly it can ruin the entire ML system
- Topics covered in this lecture:
  - Sampling techniques
  - Class imbalance problems
  - Data augmentation techniques

Sampling Techniques

- Why sampling needed?
  - Data used for training is a representative subset of real-world data.
  - Data volume is too big to process.
  - For quick experimentations.
- Two types of samplings:
  - Non-probabilistic Sampling
  - Random Sampling

Non-Probabilistic Sampling Techniques

❑ Convenience sampling.

❑ Snowball sampling.

❑ Judgmental sampling.

❑ Quota sampling.

Convenience Sampling

- A non-probabilistic sampling technique
- Subjects are selected because of their accessibility and availability to the researcher.
- A quick and easy way to gather data
- Often leads to a biased sample that may not be representative of the entire population.

Convenience Sampling

Example:

A researcher wants to study the opinions of college students about a new campus policy.

Instead of randomly selecting students from the entire student body

The researcher might choose to survey students in their own class or students they encounter in the campus cafeteria.

Snowball Sampling

Research Topic: Experiences of undocumented immigrants in a particular city.

Challenge: This population is difficult to reach through traditional sampling methods due to fear of legal repercussions, lack of formal records, and other barriers.

Snowball Sampling Process: Identify a "Seed": The researcher starts by contacting a local organization that provides services to immigrants.

Recruit Initial Participants: Through this organization, the researcher connects with a few undocumented immigrants who are willing to participate in the study.

Referral Chain: Each of these initial participants is asked to refer the researcher to other undocumented immigrants they know.

Snowball Effect: The new participants, in turn, refer others, and so on.

Snowball Sampling

Benefits of Snowball Sampling in this Case:

- **Access to a Hidden Population:** Snowball sampling allows researchers to reach individuals who might be difficult or impossible to find through other methods.
- **Efficiency:** Participants are likely to know others who share similar experiences, making it easier to recruit a large sample size.
- **Trust and Rapport:** Building trust with initial participants can facilitate deeper and more meaningful interactions with subsequent participants.

Limitations of Snowball Sampling:

- **Bias:** The sample may not be representative of the entire population, as it relies on social networks and referrals.
- **Generalizability:** Findings may not be generalizable to the broader population of undocumented immigrants.

Judgmental Sampling

- Judgmental sampling is a non-probabilistic sampling technique.
- Researcher selects participants based on their knowledge, expertise, or other relevant characteristics.
- The researcher uses their judgment to choose individuals who they believe will provide the most valuable insights for the study.

Example:

A researcher wants to study the impact of a new educational policy on high school students.

Instead of randomly selecting students, the researcher might choose to interview a small group of students who are considered to be high achievers, low achievers, and students with disabilities.

Quota Sampling

A non-probabilistic sampling technique where researchers divide the population into homogeneous subgroups (strata) and then select participants from each subgroup based on predetermined quotas

Example:

A researcher wants to study the voting preferences of a city's population. They divide the population into four subgroups based on age:

- 18-24 years old, 25-34 years old, 35-44 years old, 45 years old and above

The researcher then sets quotas for each subgroup, such as:

- 25% of the sample should be from the 18-24 age group
- 30% from the 25-34 age group
- 25% from the 35-44 age group
- 20% from the 45+ age group

Non-Probabilistic Sampling Issues

- **Lack of Representativeness:**
  - Bias: The selection process is often biased, as the researcher's judgment or convenience plays a major role.
  - Limited Generalizability: The findings from a non-probabilistic sample may not be generalizable to the larger population.
- **Uncertainty in Sampling Error:**
  - No Statistical Inference: It's difficult to calculate the margin of error or confidence intervals.
  - Limited Statistical Analysis: Many statistical techniques rely on random samplings.
- **Potential for Systematic Bias:**
  - Self-Selection Bias: Participants may volunteer due to specific motivations, leading to a biased sample.
  - Researcher Bias: The researcher's subjective choices can influence the selection of participants, potentially skewing the results.

Probabilistic Sampling Techniques

❑ Random Sampling

- Simple Random Sampling.
  - All samples in the population have equal probability to be selected.
  - Easy to implement
  - Rare categories many not appear in the training data
  - Fraud detection (population of good 99% - population of bad 1%)
  - Creating a sample of 1%, probability of bad sample in the training data is 0.0001

Sampling Techniques

❑ Random Sampling

- Simple Random Sampling.
  - All samples in the population have equal probability to be selected.
  - Easy to implement
  - Rare categories many not appear in the training data
  - Fraud detection (population of good 99% - population of bad 1%)
- Stratified Random Sampling.
  - Divide the sample into categories of interest (strata) first
  - Perform random sampling in each of the strata
  - Creating a stratified sample of 1%, probability of finding a bad sample in the training data is 0.01

Sampling Techniques

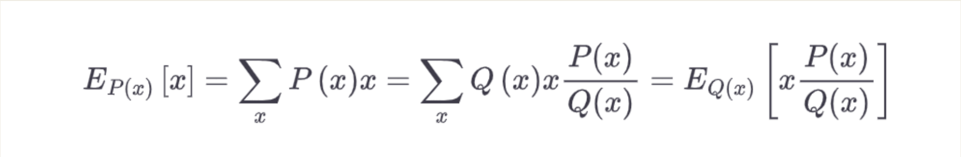
❑ Random Sampling

- Weighted Sampling.
    - Each sample is assigned a weight w.
    - The probability of being selected depends on w
- Reservoir Sampling.
    - Used for sampling from streaming data.
    - Place first k elements into an array called *reservoir*.
    - For each incoming nth element, generate a random number I between 1 and n
    - If I is between 1 and k, replace the sample in the reservoir with the nth sample, else do nothing.
    - It can be proved that each element has probability of k/n being in reservoir

Sampling Techniques

❑ Random Sampling

- Importance Sampling.
    - A method to sample from a complex distribution using a simple distribution as proxy.
    - Used extensively for Monte Carlo simulations.
    - Let the complex distribution that we want to sample x from is P(x)
    - And Q(x) is a distribution that is easy to sample from
    - Q(x) is called *Proposal Distribution or Importance Distribution*
    - Instead of sampling from P(x), sample x from Q(x) and weigh that sample by P(x)/Q(x)
    - Q(x) can be any distribution satisfying Q(x) > 0 when P(x) != 0



$$E_{P(x)}\left[x\right] = \sum_{x} P\left(x\right)x = \sum_{x} Q\left(x\right)x\frac{P(x)}{Q(x)} = E_{Q(x)}\left[x\frac{P(x)}{Q(x)}\right]$$

Sampling Techniques

❑ Random Sampling

  - Importance Sampling Demonstration in Google Colab

Class Imbalance Problem

❑ Due to substantial difference in the number of samples in each class in the training data

❑ Example – Detecting Lung Cancer from X-ray Images:

- 99.99% X-rays are from normal images
- 0.01% X-rays are from cancer patients

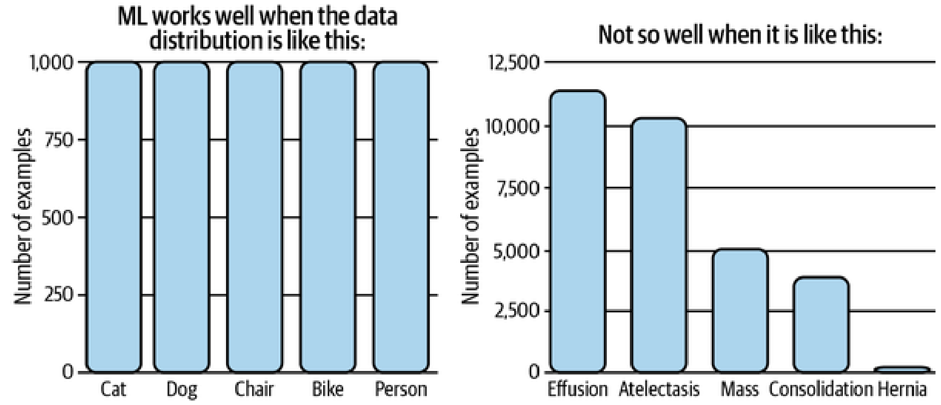


Figure 4-8. ML works well in situations where the classes are balanced. Source: Adapted from an image by Andrew Ng[26]

Class Imbalance Problem

❑ Issues with Class Imbalance:

- There is less signal for the model to learn from minority classes
- Model could get stuck in a non-optimal solution
- Cost of error estimation could be asymmetric
  - The cost of false prediction on a rare class could be higher

Class Imbalance Problem

❑ Handling Class Imbalance:

- Metrics Level – Using the right metrics
- Data Level - Resampling
- Algorithm Level – Using robust loss functions and algorithms against imbalance.

Class Imbalance Problem

❑ Handling Class Imbalance at Metric Level :

- Balanced Accuracy:
  - Arithmetic Mean of Sensitivity (True Positive Rate) and Specificity (True Negative Rate)
  - Sensitivity or True Positive Rate = TP/(TP + FN)
  - Specificity or True Negative Rate = TN(TN + FP)
  - When is Sensitivity is important ?
  - When is Specificity is important ?
  - Why simple accuracy is not a good metric ?

Class Imbalance Problem

❑ Handling Class Imbalance at Metric Level :

- F1-Score:
  - Harmonic Mean of Precision and Recall
  - 1/F1 = 1/P + 1/R
  - Precision = TP/(TP + FP).
  - Recall = TP/(TP + FN)

Class Imbalance Problem

❑ A simple example of imbalanced dataset :

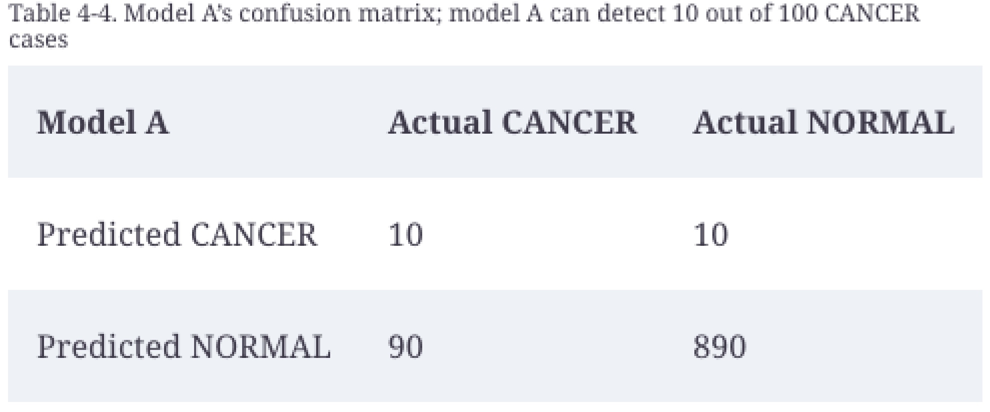


Table 4-4. Model A's confusion matrix; model A can detect 10 out of 100 CANCER cases

Model A	**Actual CANCER**	**Actual NORMAL**
Predicted CANCER	10	10
Predicted NORMAL	90	890

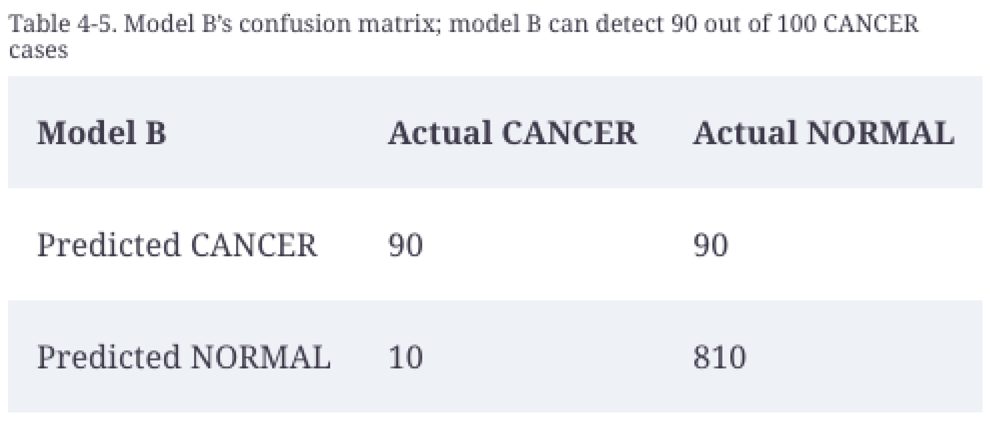


Table 4-5. Model B's confusion matrix; model B can detect 90 out of 100 CANCER cases

Model B	**Actual CANCER**	**Actual NORMAL**
Predicted CANCER	90	90
Predicted NORMAL	10	810

Class Imbalance Problem

❑ Handling Class Imbalance at Data Level :

- Resampling

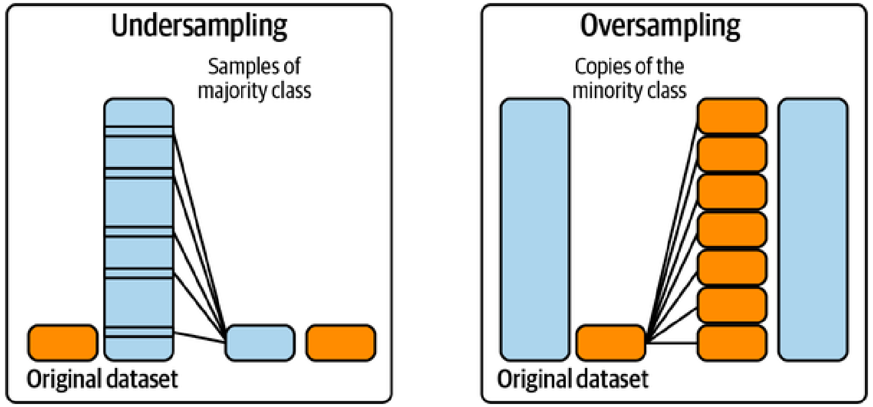


Figure 4-10. Illustrations of how undersampling and oversampling work. Source: Adapted from an image by Rafael Alencar[37]

- SMOTE – Synthetically Minority Oversampling Technique
  - Interpolate between existing minority class instances

Class Imbalance Problem

❑ How SMOTE works :

- **Identify Minority Class Samples:** The algorithm first identifies the minority class samples in the dataset.
- **Find Nearest Neighbors:** For each minority class sample, it finds its k nearest neighbors.
- **Create Synthetic Samples:**
  - *A new synthetic sample is created by taking the difference between the feature vector of the minority class sample and one of its randomly selected nearest neighbors.*
  - *This difference is multiplied by a random number between 0 and 1 and added to the feature vector of the minority class sample.*
  - *This process is repeated for each minority class sample and its k nearest neighbors, generating new synthetic samples*

Class Imbalance Problem

❑ How SMOTE works :

- A simple implementation of SMOTE for Binary Classification

Class Imbalance Problem

❑ Handling Class Imbalance at Algorithm Level :

- **Cost Sensitive Loss Function**
  - Misclassification of different classes would have different costs
  - Use a Cost Matrix in loss function to capture this

$$L\left(x;\theta\right)=\sum_j C_{ij}P\left(j|x;\theta\right)$$

- **Class Balanced Loss Function**
  - Weight of each class is inv proportional to number of samples in that class

$$L\left(x;\theta\right)=W_i\sum_j P\left(j|x;\theta\right)\mathrm{Loss}\left(x,j\right)$$

Data Augmentation Techniques

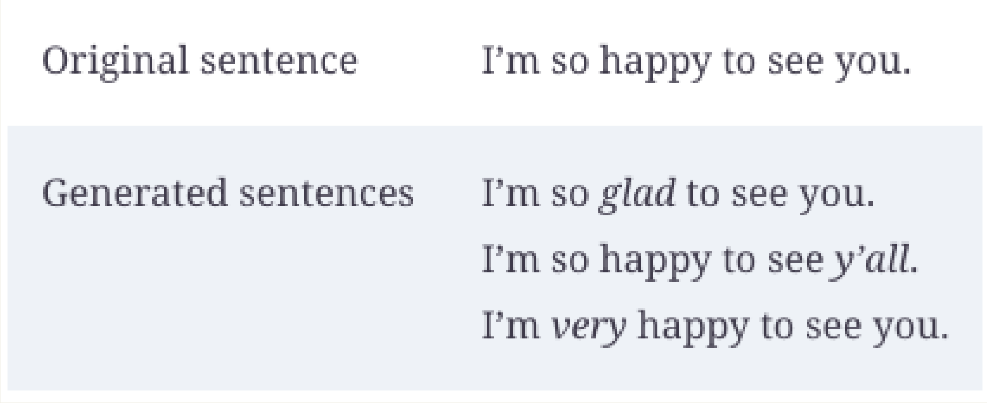
❑ Useful to increase the number of minority class samples

❑ To increase robustness of the models

❑ To prevent adversarial attacks

Data Augmentation Techniques

- Label Preserving Transformations
- For Images:
  - Crop, rotate, flip, invert, erase
- For Text:
  - Replace words with similar words



Original sentence	I'm so happy to see you.
Generated sentences	I'm so *glad* to see you. I'm so happy to see *y'all*. I'm *very* happy to see you.

Data Augmentation Techniques

❑ Perturbation

❑ Adding noise to images

❑ DNNs are sensitive to noise

❑ Can lead to misclassification