COMPSCI 589
Lecture 14: Advanced Performance Assessment

Benjamin M. Marlin

College of Information and Computer Sciences
University of Massachusetts Amherst

Slides by Benjamin M. Marlin (marlin@cs.umass.edu).
Created with support from National Science Foundation Award# IIS-1350522.

The Classification Task

Definition: The Classification Task

Given a feature vector $\mathbf{x} \in \mathbb{R}^D$ that describes an object that belongs to one of C classes from the set $\mathcal{Y}$, predict which class the object belongs to.

Error, Accuracy

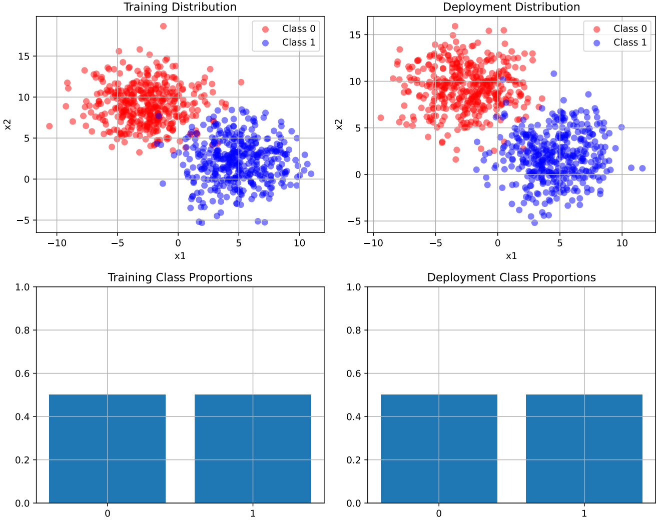
- Classification Error Rate (E, ↓): Number of incorrectly classified instances over the data set size.

$$E = \frac{1}{N}\sum_{n=1}^{N}[y_n \neq f(\mathbf{x}_n)]$$

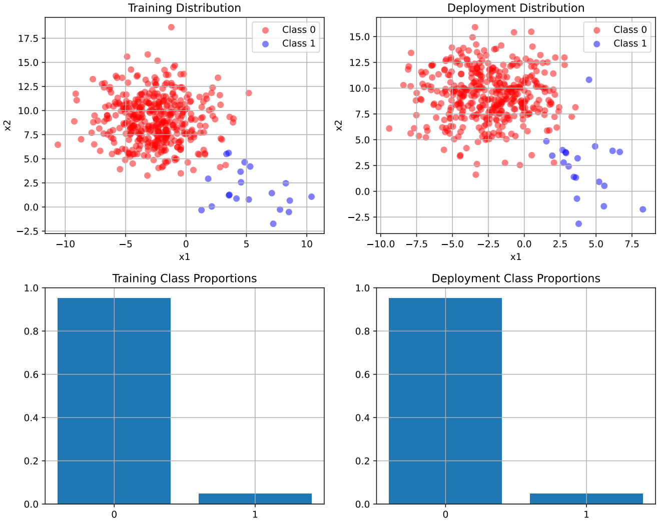
- Classification Accuracy Rate (A,↑): Number of correctly classified instances over the data set size.

$$A = \frac{1}{N}\sum_{n=1}^{N}[y_n = f(\mathbf{x}_n)]$$

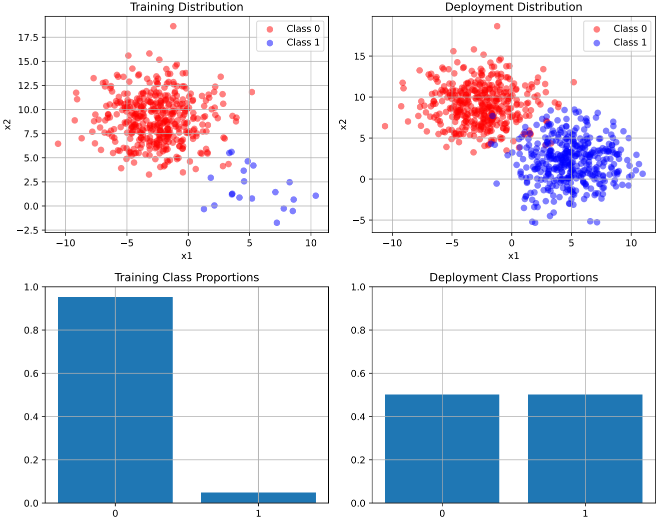
Example: Balanced Classes



Example: Imbalanced Classes, In-Distribution Deployment



Example: Imbalanced Classes, OOD Deployment



Balanced and Weighted Measures

- Balanced Error Rate (BE, ↓): Average of the per-class error rates.

$$BE = \frac{1}{C}\sum_{c=1}^{C}\frac{\sum_{n=1}^{N}[y_n = c][y_n \neq f(\mathbf{x}_n)]}{\sum_{n=1}^{N}[y_n = c]}$$

- Class-Weighted Classification Error Rate (Ew, ↓): Allows weighting errors on a per-class basis.

$$Ew = \frac{\sum_{n=1}^{N} w_{y_n}[y_n \neq f(\mathbf{x}_n)]}{\sum_{n'=1}^{N} w_{y_{n'}}}$$

Cost-Based Measures

- Misclassification Cost (MC, ↓): Avergage over test instances of the misclassification cost based on cost matrix C.

$$MC = \frac{1}{N} \sum_{n=1}^{N} C[y_n, f(\mathbf{x}_n)]$$

- Allows for different costs for each (true, predicted) pair of label values.

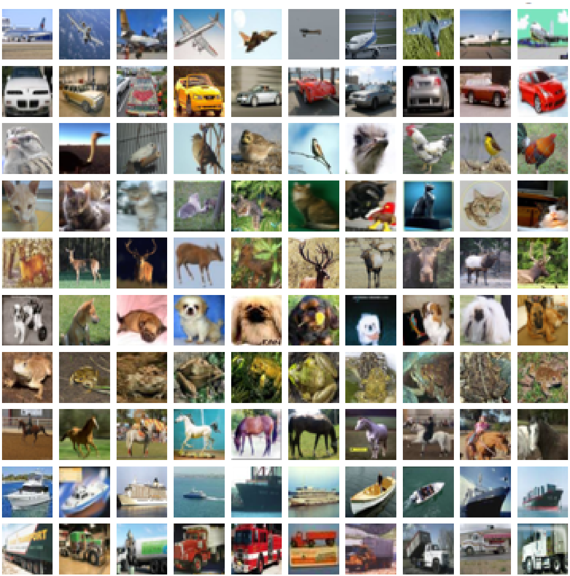
Confusion Matrix

- A confusion matrix M is a representation of the number of instances where the true class y and the class predicted by the model is y' for all pairs of labels y and y'.
- The entries in the confusion matrix are computed using:

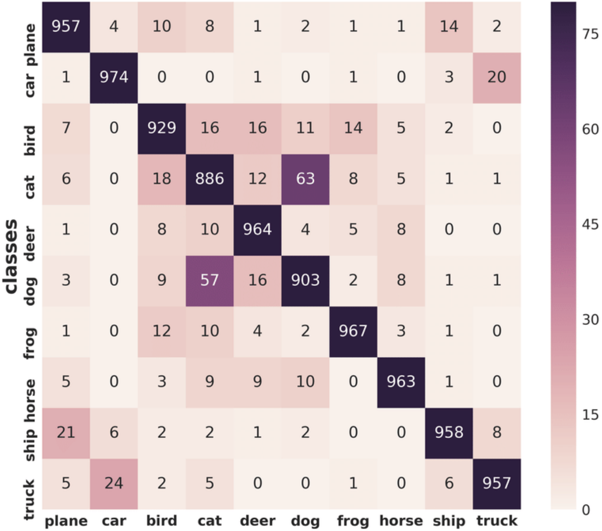
$$M_{y,y'} = \sum_{n=1}^{N} [y_n = y][f(\mathbf{x}_n) = y']$$

- The accuracy rate is equal to the sum of the diagonal entries of M divided by the number of instances.
- The error rate is equal to the sum of the off-diagonal entries of M divided by the number of instances.

Example: CIFAR10



Example: CIFAR10



Probabilistic Measures

- Negative Log Likelihood (NLL, ↓): Negative of average of log probability of test set labels.

$$NLL = -\frac{1}{N}\sum_{n=1}^{N}\log P(Y = y_n|\mathbf{X} = \mathbf{x}_n)$$

- Log Likelihood (LL, ↑): Average of log probability of test set labels.

$$LL = \frac{1}{N}\sum_{n=1}^{N}\log P(Y = y_n|\mathbf{X} = \mathbf{x}_n)$$

Weighted Probabilistic Measures

- Class-Weighted Negative Log Likelihood (NALLw, ↓): Negative of class-weighted average of log probability of test set labels.

$$NALLw = -\frac{\sum_{n=1}^{N} w_{y_n} \log P(Y = y_n | \mathbf{X} = \mathbf{x}_n)}{\sum_{n'=1}^{N} w_{y_{n'}}}$$

- Expected Misclassification Cost (EMC):

$$EMC = \frac{1}{N} \sum_{n=1}^{N} \sum_{y'} P(Y = y' | \mathbf{X} = \mathbf{x}_n) C[y_n, y']$$

Performance Measures for Binary Classification

Binary Classification Confusion Matrix

	Positive Prediction	Negative Prediction
Positive Class	# True Positives (TP)	# False Negatives (FN)
Negative Class	# False Positives (FP)	# True Negatives (TN)

Performance Measures for Binary Classification

- Precision (P, ↑): The fraction of true positives to total positive predictions.

$$P = \frac{\sum_{n=1}^{N} [y_n = 1][f(\mathbf{x}_n) = 1]}{\sum_{n=1}^{N} [f(\mathbf{x}_n) = 1]} = \frac{TP}{TP + FP}$$

- Recall (R, ↑): The fraction of true positives to total positives instances.

$$R = \frac{\sum_{n=1}^{N} [y_n = 1][f(\mathbf{x}_n) = 1]}{\sum_{n=1}^{N} [y_n = 1]} = \frac{TP}{TP + FN}$$

- F1 Score: $2(P \cdot R)/(P + R)$.

Performance Measures for Binary Classification

- True Positive Rate (TPR, $\uparrow$): The fraction of true positives to total positives instances (same as Recall).

$$TPR = \frac{\sum_{n=1}^{N} [y_n = 1][f(\mathbf{x}_n) = 1]}{\sum_{n=1}^{N} [y_n = 1]} = \frac{TP}{TP + FN}$$

- False Positive Rate (FPR, $\downarrow$): The fraction of false positives to total negative instances.

$$FPR = \frac{\sum_{n=1}^{N} [y_n = 0][f(\mathbf{x}_n) = 1]}{\sum_{n=1}^{N} [y_n \neq 1]} = \frac{FP}{FP + TN}$$

Performance Measures for Binary Classification

- Suppose we have a probabilistic binary classifier that output's $P(Y = 1|\mathbf{X} = \mathbf{x})$
- We would normally classify an instance as positive if $P(Y = 1|\mathbf{X} = \mathbf{x}) \geq 0.5$.
- However, we can achieve different tradeoffs between the true/false positives by introducing a threshold parameters τ and considering an instance to be positive if $P(Y = 1|\mathbf{X} = \mathbf{x}) \geq \tau$.
- This allows us to specify a TPR and FPR for every value of τ.

Performance Measures for Binary Classification

- True Positive Rate: The fraction of true positives to total positives instances.

$$TPR(\tau) = \frac{\sum_{n=1}^{N} [y_n = 1][P(Y = 1|\mathbf{X} = \mathbf{x}_n) \geq \tau]}{\sum_{n=1}^{N} [y_n = 1]}$$

- False Positive Rate: The fraction of false positives to total negative instances.

$$FPR(\tau) = \frac{\sum_{n=1}^{N} [y_n = 0][P(Y = 1|\mathbf{X} = \mathbf{x}_n) \geq \tau]}{\sum_{n=1}^{N} [y_n = 0]}$$

- We obtain a *receiver operating characteristic* (ROC) curve by sweeping the value of τ and plotting $TPR(\tau)$ vs $FPR(\tau)$.

Performance Measures for Classification

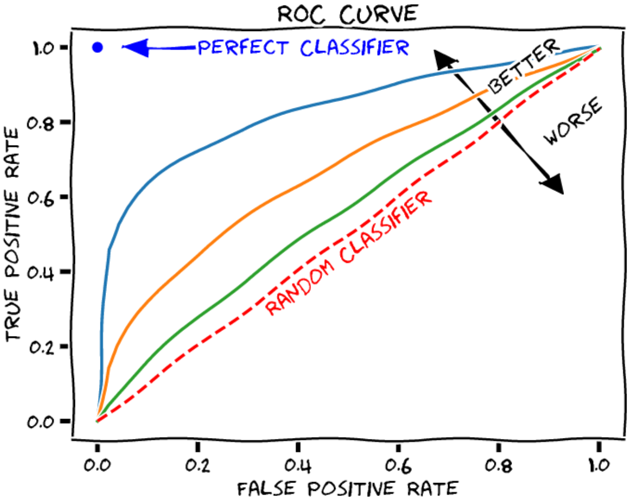


Image Credit: Martin Thoma

Performance Measures for Classification

- We can summarize an ROC curve using the area under the curve.
- This measure is referred to as AUC.
- Random guessing will yield an AUC of 0.5 regardless of class balance.
- The maximum possible AUC achieved by a classifier with a TPR of 1 and FPR of 0 is 1.
- Higher AUC values indicate better performance.

Classification Calibration Curves

Calibration curves assess how predicted probabilities align with true outcome frequencies.

- Partition the $[0, 1]$ interval into M bins:

$$I_m = \left[\frac{m-1}{M}, \frac{m}{M}\right), \quad m = 1, \ldots, M.$$

- Assign data cases to bins based on predicted probability of class 1.

$$B_m = \{n \mid 1 \leq n \leq N,\ P(Y = 1|\mathbf{X} = \mathbf{x}_n) \in I_m\}.$$

Classification Calibration Curves (Part 2)

- For each bin B_m compute the average probability of class 1:

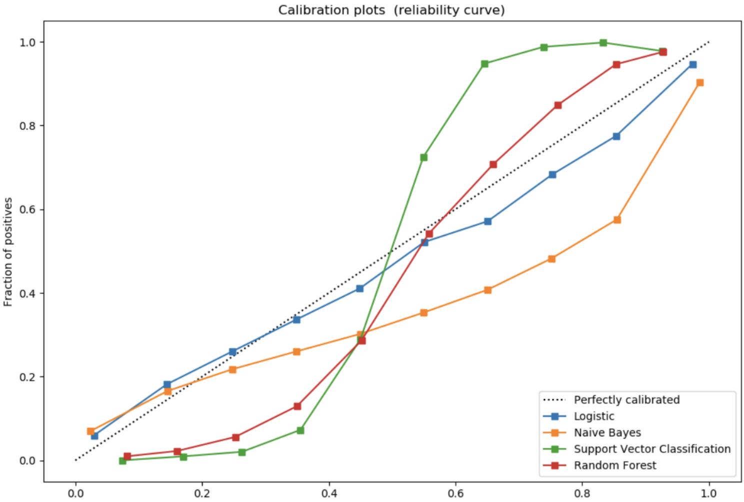
$$\text{prob}(B_m) = \frac{1}{|B_m|} \sum_{n \in B_m} P(Y = 1 | \mathbf{X} = \mathbf{x}_n)$$

- and the true frequency of class 1:

$$\text{freq}(B_m) = \frac{1}{|B_m|} \sum_{n \in B_m} \mathbf{1}\mathbb{I}(y_n = 1)$$

- Lastly, plot $\text{freq}(B_m)$ vs $\text{prob}(B_m)$.
- Perfect calibration means all points lie on a line with slope of 1.

Classification Calibration Curves



Expected Calibration Error (ECE)

- ECE (↓) quantifies overall miscalibration using weighted average absolute calibration error per bin:

$$\text{ECE} = \frac{1}{N} \sum_{m=1}^{M} |B_m| \cdot |\text{freq}(B_m) - \text{prob}(B_m)|$$

- Lower ECE means better calibration.

The Regression Task

Definition: The Regression Task

Given a feature vector $\mathbf{x} \in \mathbb{R}^D$, predict its corresponding output value $y \in \mathbb{R}$.

Metric-Based

- Mean Squared Error (↓): $MSE = \frac{1}{N}\sum_{n=1}^{N}(y_n - f(\mathbf{x}_n))^2$
- Root Mean Squared Error (↓): $RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - f(\mathbf{x}_n))^2}$
- Mean Absolute Error (↓): $MAE = \frac{1}{N}\sum_{n=1}^{N}|y_n - f(\mathbf{x}_n)|$

Statistical

- Similar to class imbalance, regression models have a scale issue when interpreting results.
- One way to fix this problem is to consider the relative performance of a model compared to a baseline approach like predicting the mean of the target values $\bar{y}$.
- The coefficient of determination, fraction of explained variation and R^2 statistic all refer to the same measure:

$$R^2 = 1 - \frac{\sum_{n=1}^{N}(y_n - f(\mathbf{x}_n))^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2}$$

- Note: This measure can be negative! Higher values indicate better performance.

Likelihood-Based

- Negative Average Log Likelihood (↓):

$$NALL = -\frac{1}{N}\sum_{n=1}^{N} \log p(Y = y_n | \mathbf{X} = \mathbf{x}_n)$$

- Average Log Likelihood (↑):

$$ALL = \frac{1}{N}\sum_{n=1}^{N} \log p(Y = y_n | \mathbf{X} = \mathbf{x}_n)$$

- Note: Since $p(Y = y_n|\mathbf{X} = \mathbf{x})$ is a probability density taking values in $\mathbb{R}^{\geq 0}$, both the NALL and ALL can be positive or negative.

Coverage

- Like with classification, probabilistic regression has a notion of calibration of uncertainty.
- Under the assumption that the conditional distribution is Normal where $\hat{y}_n$ is the mean and σ_n is the standard deviation, we can estimate the coverage of the 95% central interval using:

$$\text{coverage}(0.95) = \frac{1}{N}\sum_{n=1}^{N}\mathbb{I}(\hat{y}_n - 1.96\sigma_n \leq y_n \leq \hat{y}_n + 1.96\sigma_n)$$

- When uncertainty is correctly calibrated, $\text{coverage}(0.95)$ should be approximately 0.95.
- We can generalize this to other coverage intervals, as well as other distributions.