**AI in Drug Discovery and Development**

**Prof. Rajnish Kumar**
**Dept. of Pharmaceutical Engineering and Technology**
**IIT-(BHU), Varanasi**

**Week-03**
**Lecture-14**

Welcome to the course "AI in Drug Discovery and Development." In today's session, we will talk about evaluation metrics for AI models. So, by the end of this lecture, you will be able to understand the role of evaluation metrics in assessing AI model performance. Interpret key metrics for classification, regression, clustering, and ranking tasks. Recognize common pitfalls in metric selection and their real-world consequences. And apply best practices to select metrics that are aligned with task type, data characteristics, and decision impact.

So, in the earlier sessions, we talked about building a machine learning or artificial neural network model, as well as how we select, featurize, and pre-process the data. So, now that we have built a model, the question is how good our model is, right? So, how do we know that our model is working? So, this is, you know, the challenge of drug discovery. So, while AI is revolutionizing how we find and develop drugs, how do we quantify success in this complex field? So, evaluation metrics are methods that provide a way to measure the performance of AI models. So, it helps us understand how accurate our predictions are.

How reliable our results are and how well the model generalizes to new data. Based on different types of models, there are various types of ML models. So, there are different evaluation metrices. So, for example, for classification models, we use accuracy, precision, recall, F1 score, ROC, AUC, PR AUC, and log loss. So, these are the methods for evaluating the classification models.

And then, for regression modeling, we generally use mean absolute error, mean squared error, root mean square error, or the R-squared score. And then, for clustering models, we use the Silhouette score, Davies-Bouldin index, or the adjusted Rand index. And then, for ranking or recommendation modeling, we use the NDCG, which is normalized discounted cumulative gain, the MAP, mean average precision, or the MRR, mean reciprocal rank. So before going into those metrics, let us get to know some of these key terms, such as the true positive. So it is the case that the model predicted "yes," and the real output was also "yes."

So, an example could be an active compound that is predicted correctly as active. So when we take a training data set, we have both the active compounds and the inactive compounds

in that training data set. And if a model that we have trained on the training dataset can predict an active compound as an active compound, it is called a true positive. It means it has labeled the compound as positive or the data point as positive, and that is true as well. However, on the contrary, we have a true negative where the model has labeled it as negative, and it is negative as well.

So, it is the case where the model predicted no, and the real output is also no. So, the inactive compound was correctly predicted to be an inactive compound. However, another term is the false positive, which labels the compound as positive and the data point as positive. However, it is not true. So, in this case, the compound is inactive.

So, it is the case where the model predicted yes, but it was actually no. It has labeled an inactive compound as the active compound, resulting in a false negative; it has labeled it as negative, but that is not true. So, it was positive that it was an active compound, but the model incorrectly predicted it as inactive. So, it is the case where the model predicted no, but it was actually yes. So, these four are possible outcomes, especially when we use the classification task.

So, then we have the true positive rate which is indicating the sensitivity of the model. So, it is considered the portion of positive data points that are correctly identified as positive with respect to all data points that are positive. So, the true positive rate is the number of data points that are labeled as positive and that are actually positive, divided by the total number of data points, which is the sum of the true positives and the false negatives, meaning the total number of active compounds.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

So, the total number of predicted active compounds divided by the total number of active compounds is the true negative rate is also known as specificity. So, it is considered the portion of negative data points that are correctly identified as negative with respect to all data points that are negative. So, the total number of negatively predicted compounds and the total number of negative inactive compounds can be stated for the whole data set.

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

And then you have the false positive rate: it is a proportion of actual negatives that were incorrectly predicted as positive by the models. Where you calculate it by the false positive divided by the true negative plus the false positive.

$$\text{FPR} = \frac{FP}{TN + FP}$$

And then there is the false negative rate, which is the proportion of actual positives that were incorrectly predicted as negatives by the model. Where you have the false negative divided by the true positive plus the false negative.

$$\text{FNR} = \frac{FN}{TP + FN}$$

Okay, so let us talk about it step by step; first, let us discuss the evaluation metrics for the classification models. So, classification metrics are quantitative measures used to evaluate the performance of a classification model by assessing how accurately it assigns data points to their correct categories or classes. It simply tells us how good a model is at predicting the correct class for each data point. So it helps determine how well a model distinguishes between the different classes and guides model selection, tuning, and comparison as well. So, accuracy is one of the metrics.

So, which is extensively used for evaluating the classification models? So it measures the overall proportion of correct predictions. So, we use it when we have a balanced dataset with a roughly equal class distribution. The active and inactive individuals are equally distributed. So, in that case, we use accuracy as a metric. So, it is calculated by true positives plus the number of true negatives divided by the total of true positives, true negatives, false positives, and false negatives.

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

And precision is another metric. So, it measures how many of all the predicted positives were truly positive. So, we use it when the false positives are costly, meaning that if we cannot afford to have false positives, such as in hit identification. So, the precision indicates true positives divided by the sum of true positives and false positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

So, of all actual positives, how many did the model correctly identify? This is what recall means: how many instances the model is able to recall of all the true positives. So, we use it when false negatives are costly, as we cannot afford them, such as in the medical diagnosis of diseases. So, recall is TP divided by the sum of true positives and false negatives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

And then logarithmic loss. So, the log loss penalizes false positives and false negatives classifications. So, it usually works well with multi-class classification; whenever we use multi-class classification, we use this logarithmic loss. So, if there are n samples belonging to n classes, then the log loss is calculated using this formula.

$$\text{Logarithmic Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} \cdot \log(p_{ij})$$

Where $y_{ij}$ indicates that sample i belongs to class j and $p_{ij}$ is the probability of sample i belonging to class j, a log loss closer to 0 is better, and an equal value of 0 indicates a perfect prediction.

And then we also use the ROC AUC, which is the area under the receiver operating characteristic curve. So, the AUC is a curve plotted between the false positive rate and the true positive rate at all different data points within the range of 0 to 1. So, the greater the value of the AUC, the better the model's performance.

So, you can see here that this is the false positive rate and this is the true positive rate. So, the diagonal line represents the 50-50, meaning that half of the time the model is able to predict true positives and half of the time it is able to predict false positives. However, if our curve is like this, if it is closer to, you know, this upper triangle, upper corner. So it is actually better. It means that our model is able to identify the true positive very efficiently.

And then the F1 score, which is the harmonic mean of precision and recall, balances both concerns. So, it is used when we have imbalanced data, where both false positives and false negatives are actually issues. So, the F1 score is equal to 2 multiplied by precision times recall divided by precision plus recall.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

And then the confusion matrix is one of the matrices that is used to evaluate the performance of a classification algorithm. So, it compares the predicted classification model output with the actual ground truth values or the true labels.

So, the matrix is used to calculate key metrics like accuracy, precision, recall, and others, providing detailed insights into the types of errors the model is making. So, the key components of the confusion matrix are true positives, true negatives, false positives, which also indicate type I errors, and false negatives, which indicate type II errors. So, this is a typical example of a confusion matrix where you can see that this table shows the number of data points recognized as true positives, false negatives, true negatives, and false positives. Okay, so that was about the classification metrics. Now, let's come to the regression evaluation metrics.

So regression evaluation metrics are quantitative measurements used to assess the performance of a regression model by evaluating how accurately it predicts continuous numerical values compared to the actual observed values. So it tells us how close the model's predicted numbers are to the actual values. So they also help us understand the model's prediction error, bias and ability to capture the trends in the data. So, MAE is one of the major metrics that is used for evaluating regression models. So, it is the average distance between the predicted values and the actual values.

So, the limitation of MAE is that it does not give any idea about the direction of the error, whether we are under-predicting or over-predicting our data. So it can be mathematically represented as follows:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

So, it measures the mean squared difference between the actual and predicted values, giving higher weight to larger errors. So, in this case, it is just different that we are actually squaring this difference.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

So, when we square it, we actually give higher weight to the large errors. So, if there are large errors in any data points, they will be given more weight. So that is the purpose of squaring it. So, when do we use it? So, we use it for regression tasks, and it is sensitive to outliers and heavily penalizes large errors, making it useful when large errors are particularly undesirable. And then we have the root mean square error. So, it is the square root of the mean squared error, providing an error metric in the same units as the target variable, which makes it easier to interpret. And then we use it in the regression task, where we need to interpret the error in the same units as the target variable.

So, it is just the square root of the MSE, actually.

$$\text{RMSE} = \sqrt{\text{MSE}}$$

And then we have the R-squared score, which is also known as the coefficient of determination. So, the R-squared score measures the proportion of variance in the target variable that is explained by the model. So, it provides a measure of the model's overall goodness of fit, and we use it to evaluate the overall performance of a regression model; R-squared is represented by this formula.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

And then, coming to the clustering evaluation matrix, the clustering matrix is used to assess the quality of the clusters formed by a clustering algorithm.

Evaluating how well data points are grouped together based on similarity and how well

the clusters are separated from one another. So, these metrics help determine the effectiveness of clustering in organizing data into meaningful groups without prior knowledge of the labels. So, there are two types: one is the internal evaluation metric, which is used when two labels are unknown, assessing the cluster structure based on cohesion and separation. We can use the external evaluation metrics, which are used when two labels are known, to compare predicted clusters with the actual labels. So the internal evaluation metrics are as follows: one of the examples is the Silhouette score.

So it measures how similar a sample is to its own cluster compared to the other clusters. A high score indicates a well-defined and separated cluster, and we use it to evaluate unsupervised clustering algorithms and determine how well separated the clusters are. So it is calculated using this formula,

$$\text{Silhouette Score} = \frac{b - a}{\max(a, b)}$$

where b minus a is divided by the maximum of a and b. where a is the mean intra-cluster distance, the distance from the sample to other points in the same cluster, and b is the mean nearest-cluster distance, the distance from the sample to the nearest cluster.

And then another metric is the Davies-Bouldin Index.
So it measures the average similarity between each cluster and its most similar one, where a lower score is better, indicating well-separated and tight clusters. So it is useful for evaluating clustering tightness and separation, especially in unsupervised clustering tasks. Where it is calculated by this formula:

$$\text{DBI} = \frac{1}{N} \sum_{i=1}^{N} \max_{j \neq i} \left( \frac{S_i + S_j}{M_{i,j}} \right)$$

where 1 by n sigma i to n makes si plus sj divided by $m_{ij}$. where $s_i$ is the intra-cluster distance of cluster i, $m_{ij}$ is the inter-cluster distance between clusters i and j, and n is the number of clusters. And then we have the external evaluation matrix for the clustering models.

So, the Rand index is one of them. So, which measures the similarity between the clustering and the ground truth by considering two positive pairs and two negative pairs? So, this is when we actually have the labels for the training data. So, the score range is from 0, which indicates random clustering, to 1, which indicates perfect clustering. Let us give an example of an email. If emails are clustered into spam and non-spam, a high Rand index

indicates that the clustering aligns well with the actual labels. So, where the Rand index is equal to 2 positive plus 2 negative divided by 2 positive plus 2 negative plus false positives plus false negatives.

$$RI = \frac{TP + TN}{TP + TN + FP + FN}$$

And then we have the adjusted RAND index, which is a modified version of the RAND index that adjusts for chance clustering, making it more robust. So the score ranges from minus 1, which indicates random clustering, to 1, which represents perfect clustering, with 0 meaning a random clustering assignment. An example is clustering handwritten digits 0 to 9 from a dataset like MNIST and comparing them to the actual digit labels; a higher ARI means that the clustering is closer to the actual labels.

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}$$

And then we come to the last type of matrix, which is the ranking or recommendation matrix. So, the ranking and recommendation matrix is used to evaluate models that generate ranked lists of items or recommendations, assessing how well the system ranks relevant items higher and how closely the ranking aligns with user preferences or the ground truth.

So, these metrics help ensure that the top-ranked items are the most relevant and preferred by users. So the mean average precision (MAP) is one of the metrics. So it is used to evaluate the quality of the model's rankings. So it calculates the average precision for each query or item and then averages them over all queries. And we use it in ranking problems, such as recommendation systems, to evaluate how well the model ranks relevant items. It is calculated using this formula:

$$\text{MAP} = \frac{1}{Q} \sum_{q=1}^{Q} \text{AP}(q)$$

where MAP is equal to 1 divided by Q, the sum of sigma Q from 1 to Q, and AP Q. Where capital Q is equal to the total number of queries, and small q is the average precision for query Q. And then you have another metric, which is the normalized discounted cumulative

gain (NDCG). So, NDCG measures the quality of a ranking by giving higher relevance to items that are ranked higher on the list. So, it emphasizes ranking relevant items at the top, and we usually use it in search engines and recommendation systems where the ranking order matters.

$$\text{NDCG@k} = \frac{DCG@k}{IDCG@k}$$

And then NDCG at k is equal to DCG at k divided by IDCG at k, where DCG at k is the discounted cumulative gain at rank k and IDCG at k is the ideal discounted cumulative gain at rank k, which is the ideal ranking. And DCG at rank k is given by this formula.

$$\text{DCG@k} = \sum_{i=1}^{k} \frac{rel(i)}{\log_2(i+1)}$$

And then you have the mean reciprocal rank, which is a metric that evaluates the rank of the first relevant item in the list of search results or recommendations.

$$\text{MRR} = \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{\text{rank}_q}$$

We use it in scenarios where the position of the first relevant item is important, such as in the search engine or the recommendation system. Let us talk about some common pitfalls in metric selection.

So the first important challenge is the over-reliance on accuracy, where high accuracy can be misleading, especially in imbalanced data sets. For example, 95% accuracy occurs when 95% of the data belongs to only one class, while 5% of the data belongs to another class. So that could be one of the challenges. And cherry-picking metrics is another important issue. We are only showing favorable metrics; hide the real performance issues.

So, we always need to report a comprehensive set of everything, all those evaluation matrices, not just one, actually. Ignoring the cost of errors, treating all errors as equal can also lead to serious failures. False negative versus false positive in disease screening. And then using an inappropriate metric for the problem type, like using ROC AUC in a heavily imbalanced data set, can be less informative; PR-AUC might actually be a better one. And then static metrics focus on where metrics should evolve as the deployment environment changes, like when we observe new user behavior.

So, with that, I think we need to go for the different matrices. And let us take a look at some of the best practices for matrix selection or choosing the right matrix. So, first of all, we need to understand the type of task. So, are we going to model a classification model where we can use accuracy, precision, recall, F1 score, ROC, AUC, PR, or the regression where we are going to use the MSE, RMSE, MAE, or R-squared score? For the ranking recommendation or clustering. So, based on our task or model type, we can choose the matrix, and another important thing is that we need to align the matrix with the business or scientific goals.

So, what is our main goal? Can we afford false negatives, can we afford false positives, or how about the overall error rate? So, based on all this, we need to choose; for example, in cancer diagnosis, recall is critical; missing a positive will be worse, so recall is a very important metric in the case of diagnosing any disease, such as cancer. And then it factors in the data characteristics, such as if we have a class imbalance, so we shall prefer precision, recall, F1 score, or PR-AUC over simple accuracy, and in regression, MAE may be better than MSE. And then we need to choose the metric that matches the decision's impact as well. So if it is a high-risk decision, like medical diagnosis and fraud detection, we need more conservative evaluations, such as recall or specificity. And if it is a low-risk decision, we can afford broader metrics like accuracy or RMSE.

So, either we shall go from multiple matrices to a single matrix, or we shall not go at all. So, we always monitor a set of complementary matrix to understand different aspects of the model performance. And we always need to have an iterative approach in which we reassess matrix selection during model development and after deployment. Because real-world data drift can change the priorities as well. So, good metric selection can be determined by scientific rigor, practical relevance, as well as ethical responsibility whenever we are making an ML or AI model.

Okay, let's move on to the summary. So, evaluating AI models requires more than just high accuracy. It demands robust, context-appropriate metrics across classification, regression, clustering, and ranking tasks. So, the key challenges include misleading performance signals in imbalanced data, improper metric selection, and failure to align evaluation with real-world risks and goals. Effective model assessment depends on choosing metrics that reflect task needs, data characteristics, and decision impacts. and building reliable AI systems demand continuous validation, critical metric selection and an awareness of ethical, practical and scientific implications.

So, I have suggested some resources that you can refer to in order to learn more about this topic. And with that, thank you.