Problem 1 (Easy)

Why must data be mean-centered before applying PCA?
PCA finds directions of maximum variance around the mean of the data. If the data is not mean-centered, the first principal component may capture the mean offset instead of the true variance structure. Mean-centering ensures that PCA identifies directions that represent actual data spread rather than the location of the data in space.

Problem 2 (Medium)

PCA preserves variance but not class separability. Explain with an example.
Consider a dataset with two classes where the largest variance occurs within each class rather than between the classes. PCA will choose directions that capture this overall variance, even if those directions do not separate the classes. For example, if two elongated clusters overlap along their long axis but are separated along a direction with smaller variance, PCA will project the data along the long axis and cause the classes to overlap, reducing separability.

Problem 3 (Hard)

t-SNE shows clean clusters, but classifiers perform poorly. Explain this paradox using high-dimensional geometry.
t-SNE is designed to preserve local neighborhood relationships, not global distances or class boundaries. In high-dimensional spaces, data points can appear well-separated when projected into two dimensions, even if they are not linearly or smoothly separable in the original space. The visualization exaggerates local structure while distorting global geometry, creating visually clean clusters that do not correspond to decision boundaries usable by classifiers. As a result, classifiers trained on the original high-dimensional data may still perform poorly despite the apparent cluster separation in the t-SNE plot.