MA289 Mathematics of AI

Lesson 09 Outline — 10 March 2026

United States Military Academy, West Point
Instructor: MAJ Patrick Kuiper

Discussion Questions: Bayes' Rule, LDA, and Logistic Regression

Question 1: What role does Bayes' rule play in Linear Discriminant Analysis (LDA)?

Answer:

Bayes' rule provides the mechanism by which Linear Discriminant Analysis (LDA) converts a model of the data-generating process into a classifier. Specifically, LDA models the class-conditional distribution

$$f_k(x) = \Pr(X = x \mid Y = k),$$

which represents the probability density of observing predictor values x given that the observation belongs to class k, as well as the class prior

$$\pi_k = \Pr(Y = k),$$

which represents the probability that a randomly selected observation belongs to class k before observing any predictors.

Bayes' rule combines these quantities to compute the posterior class probability. In general, Bayes' rule can be written as

$$\Pr(Y = k \mid X = x) = \frac{\Pr(X = x \mid Y = k)\Pr(Y = k)}{\Pr(X = x)}.$$

In the context of classification with K classes, the denominator $\Pr(X = x)$ can be expressed as a sum over all classes, yielding

$$\Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

Here, X denotes the predictor vector, Y denotes the class label, $f_k(x)$ is the class-conditional density for class k, π_k is the prior probability of class k, and the summation in the denominator ensures that the posterior probabilities across all classes sum to one.

The LDA classifier assigns an observation x to the class with the largest posterior probability. Thus, Bayes' rule is essential for transforming LDA from a density estimation problem into a classification method.

—

Question 2: What assumptions does LDA make about the distribution of the predictors, and why are these assumptions important?

Answer: LDA assumes that the predictor vector $X \in \mathbb{R}^p$ follows a multivariate Gaussian distribution within each class:

$$X \mid (Y = k) \sim \mathcal{N}(\mu_k, \Sigma),$$

where each class has its own mean vector μ_k, but all classes share a common covariance matrix Σ. These assumptions are important because they lead to a closed-form expression for the posterior probabilities and result in linear decision boundaries. The shared covariance assumption causes quadratic terms in x to cancel when comparing classes, yielding discriminant functions that are linear in x.

—

Question 3: Why is Linear Discriminant Analysis considered a generative model?

Answer: LDA is considered a generative model because it explicitly models the joint distribution of the data and the class labels:

$$\Pr(X, Y) = \Pr(X \mid Y)\Pr(Y).$$

By modeling $\Pr(X \mid Y = k)$ for each class and the class prior $\Pr(Y = k)$, LDA describes how observations are generated. In principle, one could generate synthetic data by first sampling a class label from $\Pr(Y)$ and then sampling feature values from $\Pr(X \mid Y)$. Classification is then performed by applying Bayes' rule to reverse this generative process and compute $\Pr(Y \mid X)$.

—

Question 4: How does logistic regression differ from LDA from a probabilistic modeling perspective?

Answer: Logistic regression is a discriminative model because it directly models the conditional probability $\Pr(Y \mid X)$ without modeling the distribution of X. In the binary case, logistic regression assumes

$$\log\left(\frac{\Pr(Y=1 \mid X=x)}{\Pr(Y=0 \mid X=x)}\right) = \beta_0 + \beta^\top x.$$

In contrast, LDA models $\Pr(X \mid Y)$ and $\Pr(Y)$ separately and then uses Bayes' rule to compute $\Pr(Y \mid X)$. Logistic regression makes fewer assumptions about the distribution of X, whereas LDA relies on Gaussian assumptions for interpretability and stability.

—

Question 5: Why does LDA produce linear decision boundaries, and how does this relate to logistic regression?

Answer: LDA produces linear decision boundaries because it assumes a common covariance matrix Σ across all classes. When the Gaussian class-conditional densities are substituted into Bayes' rule, the resulting discriminant function for class k takes the form

$$\delta_k(x) = x^\top \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^\top \Sigma^{-1}\mu_k + \log \pi_k,$$

which is linear in x. The decision boundary between two classes is therefore defined by a linear equation. This is closely related to logistic regression, which also produces linear decision boundaries through a linear model for the log-odds. Under the LDA assumptions, the posterior probabilities from LDA and logistic regression can be very similar.

Classification Comparison Table

Concept Review Questions with Short Answers

1. **Soft-margin SVM and C:** In a soft-margin SVM, how does C control the tradeoff between a wide margin and margin violations? What happens to bias and variance as C changes?

   Answer: Large C heavily penalizes violations, leading to a narrower margin, lower bias, and higher variance. Small C allows more violations, giving a wider margin, higher bias, and lower variance.

2. **Maximum-margin idea:** What does it mean to maximize the margin, and why can this improve generalization?

   Answer: Maximizing the margin means maximizing the distance from the boundary to the closest training points. Larger margins tend to reduce variance and improve generalization.

3. **Kernel purpose:** What is the main purpose of a kernel function in SVM?

   Answer: A kernel allows SVM to fit nonlinear boundaries by implicitly mapping data into a higher-dimensional space without explicitly computing the transformation.

4. **RBF kernel and γ:** What does γ control in an RBF SVM?

   Answer: γ controls how local each point's influence is. Large γ gives very flexible, wiggly boundaries (low bias, high variance). Small γ gives smoother boundaries (high bias, low variance).

Method	**Bias–Var**	**Main Params**	n	p	**Comp.**
Logistic Reg	↑Bias ↓Var (linear)	λ or C, L1/L2	OK small–med; ↑ n helps	Handles med/high p (reg)	Fast (iterative)
LDA	↑Bias ↓Var (if Gaussian, $\Sigma_k = \Sigma$)	Priors; Σ est.	Strong small n (if assumptions)	Weak if $p \approx n$ (cov est.)	Very fast
SVM (Hard)	Low Bias; Var ↑ if noisy	– (sep. req.)	Needs separable data	Works high p (linear)	Med
SVM (Soft)	C ↑: Bias↓, Var↑	C	Good small–med n	Linear OK high p	Med
SVM (Kernel)	Flex ↑ Bias↓, Var↑	Kernel, C, γ	Med n; large n costly	Handles nonlinear; risk high p small n	High (kernel matrix)
KNN	k ↑: Bias↑, Var↓	k, metric; scale!	Needs larger n	Poor high p (curse)	Train cheap; pred costly

Table 1: Compact comparison: n = sample size, p = dimension. Arrows indicate relative increase/decrease.

5. **Feature scaling:** Why is scaling important for some models? Name two examples.

   Answer: Scaling ensures features contribute equally to distance or similarity calculations. It is especially important for KNN and SVM.

6. **KNN and k:** How does changing k affect bias and variance?

   Answer: At $k = 1$, KNN has low bias and high variance. As k increases, the boundary becomes smoother, increasing bias and reducing variance.

7. **Curse of dimensionality:** Why does KNN struggle in high dimensions?

   Answer: In high dimensions, distances between points become similar, making nearest neighbors less meaningful and reducing predictive power.

8. **Logistic regression vs SVM:** What is the key conceptual difference?

   Answer: Logistic regression models class probabilities directly, while SVM focuses on maximizing the margin. Logistic regression is generally more interpretable.

Parameter	**Where (SVC)**	**Effect if Increased**	**Bias–Variance Interpretation (Concise)**
C	All kernels	Fewer training errors; narrower margin; more support vectors can become "active"	**Variance up, bias down.** Larger C fits training data more tightly (risk of overfitting). Smaller C allows more violations (more regularization).
`kernel`	Chooses model family (e.g., `linear`, `rbf`, `poly`)	More flexible kernels can fit more complex boundaries	**Flexibility affects bias–variance.** Linear tends to higher bias / lower variance; RBF and higher-degree polynomial tend to lower bias / higher variance.
γ	RBF, Polynomial, Sigmoid	More local / sharper influence of points (RBF); stronger scaling of dot products (Poly)	**Variance up, bias down.** Large γ yields very flexible boundaries (overfitting risk). Small γ yields smoother, more global boundaries (underfitting risk).
`degree` (d)	Polynomial kernel only	More complex polynomial interactions; more curvature	**Variance up, bias down.** Higher degree increases model complexity quickly; lower degree is smoother and more biased.
`coef0` (r)	Polynomial (also Sigmoid)	In polynomial kernels, increases the influence of lower-order terms relative to pure high-order interactions	**Often bias up, variance down (but data-dependent).** Larger `coef0` typically makes the kernel behave less "purely high-degree," which can stabilize fits; very small `coef0` can emphasize high-order interactions and increase variance.
`gamma = "scale"`	Default choice for γ (RBF/Poly)	Sets γ based on data variance and number of features	**Stabilizes variance across datasets.** Helps prevent extreme γ values when features are not standardized; still may need tuning.
`gamma = "auto"`	Alternative choice for γ (RBF/Poly)	Sets γ to $1/p$ (p = number of features)	**Simpler, can miscalibrate complexity.** May be too flexible or too smooth depending on feature scaling; often less robust than `"scale"`.

Table 2: SVM (scikit-learn `SVC`) tuning parameters and their bias–variance interpretations.