Algoritmos con Función de Reducción del Error

Función de error	Tipo de datos	Complejidad de datos	Volumen de datos recomendado	Usos típicos
MSE (Mean Squared Error)	Numéricos continuos (regresión)	Baja a media	Pequeño a medio	Predicción de valores continuos (ej: precios, temperatura).
MAE (Mean Absolute Error)	Numéricos continuos	Baja a media (robusta a outliers)	Pequeño a medio	Regresiones donde hay valores atípicos (ej: series financieras).
Huber Loss	Numéricos continuos	Media (mezcla de MSE y MAE)	Medio	Regresión en presencia de outliers, más estable que MSE.
Log-Cosh Loss	Numéricos continuos	Media	Medio	Regresión suave con penalización menos agresiva que MSE.
Cross-Entropy (Categorical / Binary)	Datos categóricos (clasificación)	Media a alta	Medio a masivo	Clasificación multiclase (imágenes, texto, voz).
KL Divergence (Kullback–Leibler)	Distribuciones de probabilidad	Alta	Medio a masivo	Modelos probabilísticos, autoencoders variacionales (VAEs).
Hinge Loss	Categóricos binarios (margen de decisión)	Media	Medio	Clasificadores tipo SVM, detección de margen en redes neuronales.
Focal Loss	Datos categóricos muy desbalanceados	Alta	Masivo	Detección de objetos, clasificación con clases raras (ej: fraude).
Poisson Loss	Conteo discreto	Baja a media	Pequeño a medio	Modelos de conteo (ej: número de visitas, ocurrencias de eventos).
CTC Loss (Connectionist Temporal Classification)	Secuenciales (texto, voz)	Alta	Medio a masivo	Reconocimiento de voz, OCR, secuencias sin alineación explícita.
Triplet Loss	Datos embebidos (imágenes, texto)	Alta	Medio a masivo	Aprendizaje de representaciones (face recognition, embeddings).
Cosine Similarity Loss	Embeddings vectoriales	Alta	Masivo	NLP, recomendadores, comparación de documentos.
Wasserstein Loss	Distribuciones continuas	Muy alta	Masivo	Generative Adversarial Networks (GANs) más estables.

Algoritmos con Función de Optimización

Algoritmo de optimización	Complejidad de datos	Volumen de datos recomendado	Tipos de datos / modelos	Usos típicos
Batch Gradient Descent	Baja (datos simples, convexos)	Pequeño (costo alto en grandes datasets)	Regresión lineal / logística	Casos académicos o de prototipado.
Stochastic Gradient Descent (SGD)	Media a alta (datos ruidosos, no convexos)	Medio a masivo	Todo tipo (visión, texto, audio)	Base de la mayoría de los entrenamientos modernos.
SGD + Momentum	Media a alta (mínimos locales)	Medio a masivo	CNNs, RNNs, MLPs	Acelera convergencia y evita oscilaciones.
Nesterov Accelerated Gradient (NAG)	Alta (superficies irregulares)	Medio a masivo	Redes profundas	Mejora respecto a Momentum anticipando la actualización.
Adagrad	Media (datos esparsos)	Medio a masivo	NLP, embeddings	Ajuste adaptativo por parámetro (bueno en datos dispersos).
RMSProp	Alta (datos ruidosos, secuenciales)	Medio a masivo	RNNs, LSTM, señales	Maneja bien gradientes explosivos.
Adam (Adaptive Moment Estimation)	Alta (modelos complejos, datos no lineales)	Medio a masivo	CNNs, RNNs, Transformers	Estándar de facto en DL; rápido y robusto.
AdamW	Alta	Masivo	Transformers, NLP	Variante de Adam con mejor regularización L2.
Nadam	Alta	Masivo	Redes profundas (NLP, visión)	Combina Adam + Nesterov, mejor estabilidad.
AdaMax	Alta	Masivo	Redes muy profundas	Versión de Adam más estable en norm infinito.
L-BFGS	Baja a media (problemas suaves)	Pequeño a medio	Redes poco profundas, optimización convexa	Ej: ajustes finos en pequeños modelos.
Rprop (Resilient Propagation)	Media	Pequeño a medio	Datos numéricos tabulares	Útil cuando magnitud de gradiente varía mucho.
Lookahead Optimizer	Alta	Masivo	Grandes modelos	Hace más suave la convergencia combinando optimizadores base.
Shampoo / Adafactor	Muy alta (grandes modelos distribuidos)	Masivo (billones de parámetros)	Transformers a escala (GPT, BERT)	Optimización eficiente en memoria para training distribuido.

Algoritmos con Función de Activación

Función de activación	Complejidad de datos	Volumen de datos recomendado	Tipos de datos / modelos	Usos típicos
Step Function	Muy baja (lineal / binaria)	Pequeño	Perceptrón simple	Clases binarias, introducción académica.
Linear	Baja (lineal)	Pequeño a medio	Regresión lineal	Modelos que no requieren no linealidad.
Sigmoid (Logística)	Media (no lineal, saturación)	Pequeño a medio	Clasificación binaria, salidas probabilísticas	Probabilidades entre 0 y 1, pero con problema de gradiente desaparecido.
Tanh (Hiperbólica)	Media (centrada en 0, no lineal)	Pequeño a medio	RNNs, modelos clásicos	Mejor que Sigmoid en capas ocultas, pero también sufre gradiente desaparecido.
ReLU (Rectified Linear Unit)	Alta (simple, no saturante positivo)	Medio a masivo	CNNs, MLPs, visión por computadora	Estándar actual: rápido, evita saturación positiva.
Leaky ReLU	Alta	Medio a masivo	CNNs, redes profundas	Variante de ReLU que evita "muertes" de neuronas (gradiente = 0).
Parametric ReLU (PReLU)	Alta	Medio a masivo	CNNs, redes profundas grandes	Similar a Leaky ReLU pero con pendiente aprendida.
ELU (Exponential Linear Unit)	Alta (suaviza negativo)	Medio a masivo	Redes profundas	Mejora la convergencia, más estable que ReLU en algunos casos.
Swish (Self-Gated)	Muy alta	Masivo	Redes profundas modernas (NLP, visión)	Más suave que ReLU, propuesto por Google, mejor desempeño en grandes modelos.
Mish	Muy alta	Masivo	Modelos de visión y NLP	Nueva alternativa a ReLU/Swish con buenos resultados empíricos.
Softmax	Media	Medio a masivo	Salidas categóricas multiclase	Convierte logits en distribuciones de probabilidad.
Maxout	Muy alta (generaliza ReLU)	Medio a masivo	CNNs, modelos grandes	Representaciones más flexibles, evita saturación.
GELU (Gaussian Error Linear Unit)	Muy alta	Masivo	Transformers, NLP (BERT, GPT)	Reemplazo moderno de ReLU/Swish en LLMs.