UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

Temas avanzados en Ciencia de Datos e Inteligencia Artificial
Quiz N°1
Posgrado 2025-o

Apellidos: Méndez Lázaro Nombres: Luis Fernando

Fecha: 14/01/2025 Nota:

Indicaciones:
La Duración es de **30 minutos**.
La evaluación consta de **14 preguntas**.

1. Cuando el expected risk es muy alto debido a un underfitting severo, ¿qué componente tiende a ser dominante en el error?
   a) El error debido a alta varianza.
   b) El error debido al ruido independiente.
   c) El error debido a un sesgo alto.
   d) La falta de regularización l_2.

2. Se desea estimar un linear model mediante Maximum a Posteriori (MAP) con regularización l_1. ¿Qué distribución a priori típicamente se debe asumir?
   a) Una distribución exponencial con parámetro λ.
   b) Una distribución Gaussiana con media cero.
   c) Una distribución Beta con parámetros α y β.
   d) Una distribución Laplaciana con parámetro λ.
   e) Una distribución uniforme en la esfera unitaria.

3. ¿Cuál es la motivación principal de Kaiming/He Initialization?
   a) Reducir la complejidad computacional durante el backpropagation.
   b) Impedir la saturación de las neuronas sigmoidales en la salida.
   c) Garantizar que la salida de la función de activación sea siempre mayor que cero.
   d) Asegurar la ortogonalidad de las matrices de pesos.
   e) Mantener la varianza de las salidas de cada capa aproximadamente constante.

4. El objetivo principal de depthwise convolution en MobilNet es:
   a) Reemplazar pooling con convoluciones más profundas.
   b) Separar la convolución en una parte espacial y otra de combinación de canales.
   c) Reducir la dimensionalidad antes de aplicar convoluciones más grandes para reducir el costo computacional.
   d) Agregar una etapa de channel attention para mejorar la capacidad de la red.

5. El cell state en las LSTM son responsables de:
   a) Almacenar el último estado de activación para la salida de la LSTM.
   b) Capturar información de largo plazo.
   c) Ajustar dinámicamente la dimensionalidad de entrada en cada paso.
   d) Eliminar el problema de vanishing gradient al saturar las activaciones.
   e) Almacenar la información más relevante para la tarea actual de la red.

6. ¿Cuál es el objetivo de emplear 1×1 convolutions dentro de los bloques de GoogLeNet (Inception v1)?
   a) Reducir la dimensionalidad antes de aplicar convoluciones más grandes.
   b) Mantener un **receptive field** constante en cada convolución.
   c) Simplificar el entrenamiento al eliminar operaciones de pooling.
   d) Forzar la ortogonalidad entre los canales de salida.

7. En un experimento de computer vision, se entrenan dos modelos:
   - un Convolutional Neural Network (CNN) con stride mayor a 1 y
   - un Multilayer Perceptron (MLP),

   ambos con un número similar de parámetros. Partiendo de la manifold hypothesis, ¿que justifica que la CNN supera significativamente al MLP en generalización?
   a) El MLP no sufre el curse of dimensionality, pero carece de regularización para competir con la CNN.
   b) El stride mayor a 1 en la CNN multiplica la capacidad paramétrica, superando la efectividad del MLP.
   c) El MLP tiene un strong prior sobre los features, mientras que la CNN ignora la correlación espacial.
   d) La CNN, al usar convoluciones, aprovecha la estructura local y reduce la dimensionalidad efectiva al proyectar los datos en un manifold más simple.
   e) La CNN elimina por completo el problema de vanishing gradients, mientras que el MLP no puede.

8. Durante el diseño de una CNN que será parte de un modelo de segmentación de imágenes médicas, se está decidiendo el tamaño adecuado para el kernel: usar filtros 3×3 o 7×7, asumiendo que la cantidad de parámetros totales son equivalentes (ajustando la profundidad de cada capa). El objetivo de la arquitectura es capturar tanto detalles locales finos como regiones más amplias. ¿Qué argumento a favor tiene la elección de varios filtros 3×3 consecutivos en lugar de un solo filtro 7×7, asumiendo mismo costo computacional?
   a) Múltiples 3×3 generan mayor profundidad efectiva, permitiendo aprender representaciones más complejas.
   b) Con filtros 3×3 la red puede enfocarse en detalles pequeños de la entrada, permitiendo una segmentación más fina.
   c) Un 7×7 reduce la probabilidad de overfitting, porque combina más valores en un paso.
   d) 3×3 obliga a la red a ignorar la correlación espacial lejana, lo cual mejora la detección de bordes.
   e) Un solo 7×7 puede capturar detalles locales y globales simultáneamente, mientras que múltiples 3×3 no.

9. En un model de traducción automática basado en LSTM, se detecta que ciertas unidades descartan el contenido de nuevas entradas en cada timestep y nunca ingresa nueva información. ¿Cuál de los siguientes pasos sería el más adecuado para detectar la causa del problema?
   a) Obligar los gates tener siempre el valor de 0.5, asegurando equilibrio entre olvido e incorporación de nueva información.
   b) Disminuir la dimensionalidad del hidden state para forzar una mayor utilización de los gates.
   c) Revisar la distribución de los pesos de la forget gate para ver si hay saturación en la sigmoide.
   d) Eliminar la función de activación en el cell state, convirtiendo la salida en una simple multiplicación.
   e) Usar un prior Laplace en los gates, forzando sparsity lo cual mejoraría la retención de nueva información.

$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$

↑ poder de descarte

$f_t = \sigma(U \dots)$

bacth norna

10. El Zorro implementa su primera red neuronal profunda en Rust. Por falta de tiempo no llegó a implementar Batch Normalization (BN) ni inicialización de pesos. Entonces, el entrenamiento del modelo tendió a atascarse debido al vanishing gradient. Ante estos resultados, El Zorro planea implementar BN en todas las capas convolucionales, pero no ha planeado implementar la inicialización de pesos. ¿Cuál es el efecto más probable que tenga BN en este escenario?
    a) BN por sí sola asegura que la red no llega a dead ReLU, eliminando completamente el problema de vanishing gradient.
    b) La normalización de activaciones estabiliza el entrenamiento y facilita la propagación de gradientes, pero una mala inicialización de pesos aún puede afectar el rendimiento.
    c) BN empeora el problema de vanishing gradient porque crea dependencia entre muestras del mismo batch.
    d) Con BN, la función de pérdida se vuelve no convexa, generando múltiples mínimos globales.
    e) BN solo es efectiva en la capa de salida, pero no en capas intermedias.

11. En un módulo de CNN para segmentación de alta resolución, se introduce un dilated convolution con dilatación $d = 4$. Además, en capas subsiguientes se emplean convoluciones "normales" (dilatación $d = 1$). Se observa que el receptive field crece, pero ¿la densidad de muestreo espacial se verá afectada?
    a) El receptive field se reduce, ya que la dilatación salta píxeles. ?
    b) El muestreo de la imagen se vuelve uniforme al combinar convolución dilatada y no dilatada, sin aumento en la resolución.
    c) El stride y la dilatación se compensan mutuamente, resultando en un receptive field idéntico al de una convolución normal.
    d) Al usar dilatación solo en una capa, las demás capas convolucionales ignoran completamente esos saltos.
    e) Con dilated convolution, se incrementa el covertura espacial, pero pueden surgir huecos en la percepción local, lo que puede impactar la continuidad de los bordes en la segmentación.

12. Se realiza un experimento con redes de distinto tamaño. En cierto punto, se observa que una red pequeña obtiene un error elevado (underfitting), luego otra red mediana cae en overfitting y finalmente una red gigante vuelve a bajar el error en test. ¿Por qué la red gigante obtiene buena generalización a pesar de ajustarse casi perfectamente a los datos de entrenamiento?
    a) Al ser masiva, la red prescinde del uso de optimizadores estocásticos y, por ende, no overfittea.
    b) En redes con muchos parámetros no hay posibilidad de obtener error cero en entrenamiento.
    c) El cross-entropy loss induce una regularización efectiva que disminuye con el tamaño del modelo.
    d) La capacidad extremadamente grande permite una interpolación benigna que, incluso con overfitting perfecto, generaliza en el test.
    e) La mediana red entra en un régimen saturado de ReLU que la masiva evita completamente por la inicialización.

13. Suponga que tienes un dataset con miles de atributos irrelevantes y solo unas pocas variables relevantes. Estudiamos dos modelos:
    - Linear model con regularización l_2 y
    - Linear model con regularización l_1.

    Ambos alcanzan similar empirical risk en entrenamiento. Sin embargo, en validación, el modelo con l_1 logra un menor error. ¿Cuál sería la causa?
    a) La regularización l_1 incrementa la magnitud los coeficientes relevantes sin penalizar los irrelevantes.
    b) La regularización l_2 colapsa los coeficientes a valores idénticos, lo que destruye la capacidad de clasificación.
    c) La regularización l_1 tiende a forzar esparsidad, apagando los atributos irrelevantes, por lo que es robusta a valores atípico.
    d) La regularización l_1 funciona mejor solo en datos pequeños, no en miles de atributos.
    e) La regularización l_2 llega overfitting en presencia de variables altamente correlacionadas, empeorando la varianza del modelo.

14. Se entrena una CNN con AdamW, que introduce un término de weight decay desacoplado de la actualización de Adam. Soldi insiste en que es equivalente a usar una regularización l_2 clásica. ¿Por qué es erronea esta afirmación?
    a) En AdamW, el weight decay se aplica a la velocidad de actualización, lo que difiere de la forma en que se emplea la regularización l_2.
    b) AdamW añade la penalización sobre los gradientes en cada actualización, mientras l_2 actúa solo en el loss.
    c) Con AdamW, la norma del gradiente no se ve afectada, mientras que en l $_2$, la norma aumenta progresivamente.
    d) La diferencia principal es la inicialización: AdamW emplea pesos muy grandes para el weight decay.
    e) Con l_2 se ajustan únicamente capas convolucionales, mientras AdamW actúa en capas densas.

Cada pregunta equivale a 1.5pts