Diseño / Hiperparámetros

¿Qué?

- Preproceso
- Ruido artificial/ampliación de muestra
- Parte por azar y parte estratificada
- Tamaño de muestra
- Número de procesadores
- Función no-lineal entre capas y a la salida
- Medida de error a optimizar
- Inicialización de pesos
- Algoritmo de ajuste
- Velocidad de ajuste. Evolución.
- Número de muestras para calcular gradiente
- Poda posterior de la red

Preproceso

- Más probables: `mediavar`, `medianavar`, `rango1`, `pca`. Mira la distribución de las variables originales:
  - Más o menos acampanada: más a `mediavar`
  - Cola gorda de valores extremos: más a `medianavar`. Si tienes variaciones enormes, quizá logarítmos
  - Muy agrupados: más a `rango1`
  - Variables relacionadas una con otra: más a `pca`
- No tanto: `unifor` Solo si una zona está demasiado pegada, pero no todo, y queremos más contraste en esa zona. Sale en problemas de imagen.
- Poco probable: `aleat`
- Por conocimiento previo, primeros principios, podemos considerar otras transformaciones

Ruido / ampliación de muestra

La ampliación artificial solo tiene sentido si vemos que necesitamos más puntos. A usar si vemos que el tamaño de muestra nos lleva a mejorar.
Ruido: centésimas o milésimas, incluyendo el 0 En general, la probabilidad decrece desde el 0 en adelante

Selección

La selección estratificada no es práctica en dimensiones muy altas, que son comunes. Si se usan ambas, la parte por estratificación aumenta con el agrupamiento de datos que detectemos.

Tamaños de muestra

La probabilidad aumenta con el tamaño. Si excedemos la muestra original, nos vamos a ampliación de muestra y eso decrecería. El máximo se daría en el entorno de la siguiente proporción:

ajuste: 70%

validación: 15%

prueba: 15%

El conjunto de prueba no está sujeto a validación, así que se dejaría fijo. Lo que puede variarse es el resto. Ninguno puede aproximarse a 0

Número de procesadores

En redes de conexión completa, salvo conocimiento previo, la probabilidad sube hasta alrededor de cinco y luego se mantiene uniforme, decreciendo suavemente hacia el infinito.

Función no-lineal

Entre capas, las más probables son las sigmoides (incluye la tangente hiperbólica) y las lineales rectificadas (incluye la variante Leaky)
A la salida, si es un problema de regresión, la más probable es la identidad, si es clasificación binaria la sigmoide, y si es de clasificación múltiple, la Softmax

Función de error

Estamos hablando de la función a optimizar en el ajuste. Para la evaluación del modelo, la función debe ser la más convieniente para medir su eficacia en uso, lo cual no se variará.
Para el ajuste, si es regresión, la más probable es el error cuadrático. Si es clasificación binaria, la entropía cruzada binaria. Si es clasificación múltiple, la entropía cruzada.

Inicialización de pesos

La por defecto (`inibase`) es ligeramente más probable

Algoritmo de ajuste

Probabilidades muy parejas. Ligeramente por encima: SGD, Adam y Rprop

Velocidad de ajuste

El valor de salida es más probable cuanto mayor, pero no demasiado. Sin conocer el caso, podría decirse que la probabilidad suba notablemente hasta la milésima, siga subiendo más suavemente hasta la décima y luego decrezca.
La evolución en el tiempo podemos separarla en las iteraciones iniciales y las siguientes.
Para las iniciales es equiprobable un aumento o una disminución con el paso (eso incluye permanecer constante). Las variaciones de tipo cuadrático serían claramente menos probables.
Para las siguientes es más probable que tienda a descender con el inverso del paso, con descenso suave hacia mantenerse constante y con descenso fuerte a un descenso con el inverso del cuadrado del paso.

Muestra para cálculo de gradiente

La probabilidad sería descendente con el tamaño, sin anularse para toda la muestra.

Poda posterior

Una poda en procesadores equivale a probar otra red con un número de unidades menor, así que no la vamos a considerar. Nos queda la poda de pesos, o conexiones. Puede ser de probabilidad constante hasta cifras del 20% y anulándose en el entorno del 50%, ya que probablemente estaríamos hablando de quitar procesadores. Pero eso si tienes una red grande. Si es pequeña, la probabilidad es siempre decreciente desde cero.

Estrategia manual - Fase inicial

1. Elige los hiperparámetros a analizar. Puedes dejar algunos de los anteriores fijos por ser demasiado complejos o porque para ese caso no los creas importantes.
2. Elige su distribución a priori. Puedes inspirarte en las observaciones anteriores. Puede ser una distribución conjunta o tomarlos independientes. En principio puedes suponerlos independientes y al ir avanzando, considerarlos conjuntamente.
3. Toma un conjunto inicial de casos, alrededor de media docena.
4. Haz con ellos un ajuste completo, salvo que manifiestamente se vea que va mal. Quédate con las curvas de ajuste. Considera especialmente las de los mejores (mejor generalización). Establece una relación entre curva de ajuste y resultado final.

Estrategia manual - Fase iterativa

1. Apunta a zona de hiperparámetros que parece mejor y toma otra muestra
2. Inicia el ajuste. Si va mal, según las curvas de ajuste guardadas, córtalo, y sustituye el resultado final por otro basado en la tendencia.
3. Si ya son demasiados casos, corta y quédate con el mejor
4. Actualiza el lote de curvas de ajuste y la relación de objetivo con hiperparámetros. Vuelve al paso 1.