Ajuste y generalización

Que aprenda y que ese aprendizaje sea robusto frente a situaciones nuevas

Fuentes de error

Limitaciones de la red

¿Cómo lo medimos?

Método de ajuste

Impredecibilidad del fenómeno

Limitaciones de la red

- Por ser red neuronal
- Por tener cierto número de capas y ciertas conexiones
- Por la cantidad de procesadores
- Por el cálculo realizado en cada uno
- Soluciones: selección de modelos
- Aproximación si es una capa oculta, ¿cuántos componentes principales representan el 70% de la varianza? En muchos casos normales, cerca de 5

¿Cómo medimos el error?

- Por propiedades teóricas, muchas veces el error cuadrático
  - Simetría de signo
  - Crecimiento parabólico: poca importancia de errores pequeños, crecimiento fuerte no acotado
- Otras posibilidades: absoluto (igual crecimiento desde el principio), Tukey (satura), ...
- En clasificación, suele usarse el error entrópico
- Por conveniencia del algoritmo de ajuste: exponenciales
- Por conveniencia de la robustez: añadir parámetros del propio modelo

¿Pardon me?

Descripción de otras medidas de error

- Entropía cruzada simétrica (para caso binario):

$$\frac{-1}{N}\sum_{1}^{N}\sum_{1}^{M}\left(y_{ij}\ln\left(r_{ij}\right)+\left(1-y_{ij}\right)\ln\left(1-r_{ij}\right)\right)=\frac{-1}{N}\left(\sum_{y_{ij}=1}\ln\left(r_{ij}\right)+\sum_{y_{ij}=0}\ln\left(1-r_{ij}\right)\right)$$

- Área bajo ROC (clasificación binaria)
  - Las respuestas están entre 0 y 1 (no/sí)
  - Se va variando el umbral de decisión no-sí
  - Se obtiene una curva poniendo en abscisas la fracción de falsos positivos y en ordenadas la de positivos acertados
  - Se mide el área bajo esa curva

Hazlo para el modelo de la pizarra

Descripción de otras medidas de error

- Coeficiente de correlación

$$\rho_s^2 = \frac{\left[\sum_{i=1}^{N}{}_s (r_i - \bar{r})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{N}{}_s (r_i - \bar{r})^2 \sum_{i=1}^{N}{}_s (y_i - \bar{y})^2}$$

- En clasificación
  - Sensibilidad **a una clase**: $\frac{RealesCorrectos}{Reales}$
  - Especificidad **de una clase**: $\frac{FalsosCorrectos}{Falsos}$
  - Valor predictivo positivo **en una clase**: $\frac{RealesCorrectos}{RealesCorrectos + FalsosFallados}$
  - Precisión: $\frac{Aciertos}{Total}$

Otra vez, para el de la pizarra

Métodos de ajuste

- Punto de partida
  - Aleatorio que no condicione mal los elementos utilizados en el algoritmo: puede que cada vez vaya a sitio distinto
  - Sacado de otro método: autocodificación
- Algoritmo de búsqueda/actualización. Puede influirle las características de la muestra
- Criterio de parada
  - Agotamiento, ¿?
  - Estabilización, ¿?

¿De dónde podemos sacar un criterio para decidir que ya lo hace suficientemente bien?
¿Qué podemos mirar para decidir que ya no va a progresar más?

Algoritmo de actualización

- Búsqueda pura
  - Algoritmos genéticos
  - Cristalización
  - “Extremo”: capa oculta grande aleatoria, sólo se ajusta la de salida
- Uso de derivadas
  - Descenso gradiente
  - Gradiente conjugado
  - Cuasi-Newton, Métrica variable, Levenberg-Marquardt

Muestra tus conocimientos de optimización, por favor, a partir de la documentación de Pytorch u otras fuentes

Problemas

- Espacio de enorme dimensión
  - Grave problema especialmente para búsqueda pura
- Mínimos locales
  - Grave especialmente con las derivadas
  - La exponencial del error cuadrático tiene menos mínimos locales
  - Evitar regiones sin gradiente: función de activación que tenga fluctuaciones en zonas “planas” (Ej: RRELU)
  - Un punto inicial distinto puede evitar ese mínimo local

¿Qué dimensión tiene el espacio de búsqueda del perceptrón de la pizarra?

Paso de avance

- ~Constante proporcionalidad a gradiente
  - Grande: rápido e inestable
  - Pequeño: lento, pero seguro
  - Probar: aumentando si se puede
  - Ciclo: empieza bajo, sube, baja

Impredecibilidad del problema

- Problemas
  1) Aleatoriedad intrínseca
  2) Tipo de función
  3) Tamaño de la muestra, representatividad⇒extrapolación, generalización
- Soluciones
  1) ...
  2) Adecuación de la red. Cálculos de procesador
  3) Diseño previo. Aumento de muestra. Control estadístico. Remuestreo.

¿Cuál es ...
...por el problema en sí?
...por cómo cogemos nosotros los datos del problema?

Muestra

- Problemas
  - Ruido, aleatoriedad
  - Cobertura, dominio
  - ¿Están todas las variables?
  - Representatividad de la distribución. Problema de las clases importantes mal representadas.
- La dependencia de la muestra crece con el número de pesos (¿ecuaciones>incógnitas?)
- Soluciones (nunca total, hay fórmulas de estimación de cota)
  - Nº de pesos (grandes):
    - error
    - especializar pesos, que no jueguen todos siempre (nº efectivo menor)
  - Oscilaciones fuertes: error
  - Regresión bayesiana
  - Muestrear la influencia de la muestra: media (es la red), varianza (es la muestra)
  - Muestra artificial ¿?: interpolación, estimación probabilista

Programa un aumento artificial de la muestra de casas

Muestra-Tamaño

- Más muestras: mejor informados, mejor aprendizaje, ¿o no?
  - ¿Cabe toda esa muestra?
  - ¿Cómo va a ir de rápido si tiene que pasar por cuatrillones de casos?
  - ¿Son todos esos casos igual de informativos?
  - ¿Todos los casos le aprovechan a la red por igual?
- Estrategia de ajuste: calcular el gradiente sobre subconjunto de muestra (minibatch)
- ¿Cómo lo elegimos?
  - Al azar
  - Por orden secuencial
  - Empezando por los casos más “fáciles” y dejando los “difíciles” para el final

Enfoque probabilista bayesiano

Buscamos un buen modelo, que sea lo más probablemente posible el real

Hipótesis

- La red aproxima la forma funcional perfectamente
- Los conjuntos de pesos distintos producen modelos distintos
- La parte no predecible tiene media nula, varianza constante y no está autocorrelada, ni correlada con las variables
- Esta parte tiene una estructura de probabilidad
- La muestra es representativa de la población total

Valora de 0 a 100 la fiabilidad de cada hipótesis

Regresión bayesiana

Max(Prob(pesos/muestra))=Max(Prob(muestra/pesos)Prob(pesos)/Prob(muestra))

Max(log(Prob(pesos/muestra)))$\Longleftrightarrow$

Max(log(Prob(muestra/pesos)+log(Prob(pesos)))

Revisa los paréntesis

Regresión bayesiana

$$\log\left(Prob\left(muestra/pesos\right)\right)=\ldots$$

$$\log\left(\prod_i \frac{e^{\frac{-(r_i-y_i)^2}{2\sigma_d^2}}}{\sqrt{2\pi\sigma_d}}\right)=\ldots$$

$$\sum_i\left(\log\frac{e^{\frac{-(r_i-y_i)^2}{2\sigma_d^2}}}{\sqrt{2\pi\sigma_d}}\right)=\ldots$$

$$-\left(N_d\log\sqrt{2\pi\sigma_d}+\sum_i\frac{(r_i-y_i)^2}{2\sigma_d^2}\right)=\ldots$$

$$-\left(Cte+\frac{1}{2\sigma_d^2}\sum_i\left(r_i-y_i\right)^2\right)$$

¿Por qué nos pulimos el valor concreto de la constante?

Regresión bayesiana

$$\log\left(Prob\left(pesos\right)\right)=\ldots$$

$$\log\left(\prod_j \frac{e^{\frac{-p_j^2}{2\sigma_p^2}}}{\sqrt{2\pi\sigma_p}}\right)=\ldots$$

$$-\left(N_p \log\sqrt{2\pi\sigma_p}+\sum_j \frac{p_j^2}{2\sigma_p^2}\right)=\ldots$$

$$-\left(K+\frac{1}{2\sigma_p^2}\sum_j p_j^2\right)$$

Repite con una función de densidad que creas mejor

Regresión bayesiana

Max(Prob(pesos/muestra))

$$Max\left(-\left(Cte+\frac{1}{2\sigma_d^2}\sum_i\left(r_i-y_i\right)^2\right)-\left(K+\frac{1}{2\sigma_p^2}\sum_j p_j^2\right)\right)$$

$$Min\left(Cte+\frac{1}{2\sigma_d^2}\sum_i\left(r_i-y_i\right)^2+K+\frac{1}{2\sigma_p^2}\sum_j p_j^2\right)$$

$$Min\left(\sum_i\left(r_i-y_i\right)^2+\frac{\sigma_d^2}{\sigma_p^2}\sum_j p_j^2\right)$$

¿Y con otras funciones de densidad?