Capítulo 1

Perceptron

1.1 Perceptron

O Perceptron foi introduzido por F. Rosenblatt na década de cinquenta. A rede neural artificial mais simples é o Perceptron de camada única esta tem vários nomes que refletem sua simplicidade e naturlidade. O Neurônio de Macullogh e Pitts ou o discriminador linear. É fundamental entender suas propriedades. Alguns autores, Minsky e Papert de forma mais contundente, escreveram sobre ele para depois deixá-lo de lado devido a limitações óbvias. Sem perceber que o futuro desenvolvimento da área deixaria seus argumentos sem força, o descartaram e nisso devolveram a énfase em inteligencia artificial à área de sistemas peritos. A importância do Perceptron se deve a vários motivos. Do ponto de vista pragmático, pode ser visto como um bloco na construção de máquinas mais poderosas. Isto inclui Máquinas de Vetor de Suporte (SVM) e Máquinas do Estado Líquido (LSM). Seu estudo em si pode ser revelador de várias propriedades que seriam impossíveis de ver ao atacar sistemas mais complicados sem o beneficio da lição de casa prévia. Ainda assim são ricos o suficiente para despertar o interesse do ponto de vista teórico e experimental. Ainda mais, o uso de algoritmos de treinamento inteligentes permite competir com SVM em eficiência.

1.2 O problema de aprendizagem supervisionado

Considere um conjunto de vetores $\xi_\mu \in \mathbb{R}^N$, cada um associado a um rótulo σ_μ que formam o conjunto de aprendizagem $\mathcal{L} = \{\xi_\mu, \sigma_\mu\}$. Esta é a informação disponível. O caso mais simples possível é quando os rótulos σ_μ tomam valores ± 1. O perceptron é definido por um vetor de pesos $\mathbf{J}$ e uma função de transferência, que pode ser uma função contínua ou não, mas aqui olharemos primeiro para funções degrau: $sign(x)$ é o sinal de x. O perceptron classifica um vetor numa classe:

$$\sigma = sign(\mathbf{J}.\xi). \tag{1.1}$$

Note que a classificação não depende dos módulos dos vetores. O problema de aprendizagem é determinar o conjunto de pesos **J** que classifique os vetores de forma adequada, medida pelo erro de treinamento e_T ou o erro de generalização e_G ou ainda o erro de predição e_P. e_T mede a memorização dos rótulos. Os rótulos podem ter sido corrompidos por algum processo de ruído e o erro de predição mede a probabilidade de que um novo vetor seja classificado de forma igual ao rótulo enquanto o de generalização mede a probabilidade de que a classificação seja correta. Estas duas medidas coincidem quando não hã ruído. O erro de generalização tem mais utilidade teórica, pois diz respeito 'a concordância com uma regra desconhecida e por inferir. O erro de predição pode ser estimado na prática.

O produto escalar determina a geometria do problema. O perceptron só pode classificar corretamente o conjunto de treinamento se houver um hiperplano que separa as duas classes, isto é só pode representar o problema de separação linear. O clássico problema do XOR, ou exclusivo, é o exemplo clássico:

Exercício Em duas dimensões, $\xi = (1, 1)$ e $(-1, -1)$ são classificados em uma categoria, e.g. $\sigma = 1$; e $\xi = (1, -1)$ e $(-1, 1)$ na outra $\sigma = -1$. Faça o desenho e convença-se que não ha reta que separe os dois primeiros e vetores dos outros dois. Mas há duas soluções para este problema. Note que podemos usar duas retas para separá-los. Usamos 3 perceptrons. Um resolve o problema de separação com associado a uma reta, o segundo o problema associado 'a outra reta. O terceiro usa como entrada a saida dos perceptrons 1 e 2 e resolve o problema. Uma segunda solução passa por aumentar a dimensão do problema. Discuta.

1.2.1 Teorema de convergência do Perceptron

Provaremos a seguir o teorema de Rosenblatt para a convergência de um algoritmo de aprendizagem. Se existe uma solução, isto é se há um vetor **B** que separa um conjunto de vetores nas categorias corretas com uma certa margem κ, então em um número finito de passos é possível encontrar algum vetor que também consiga essa separação. Chamamos acerto (ou erro) nesta seção, quando $\mathbf{J}(t).\xi_\mu\,\sigma_\mu < \sqrt{N}\kappa$(caso contrário).

O algoritmo Perceptron de Rosenblatt é muito importante e simples. Os vetores do conjunto de treinamento serão considerados sequencialmente e apresentados ao perceptron, se a classificação é correta nada é feito. Se ocorre um erro, uma mudança no vetor de pesos $\mathbf{J}(t)$ é necessária. O processo se repete com o próximo exemplo da lista até conseguir convergência. Começamos da *tabula rasa*: **J**= 0 e a correção é dar um passo paralelo ao vetor apresentado. Podemos considerar um indice temporal e pensar na dinâmica de aprendizado

$$\mathbf{J}(t+1) = \mathbf{J}(t) + \frac{f_\mu}{\sqrt{N}}\sigma_\mu\xi_\mu \tag{1.2}$$

onde f_μ, que chamaremos de função de modulação é 0 se $\mathbf{J}(t).\xi_\mu\,\sigma_\mu > \sqrt{N}\kappa$ (acerto) e 1 caso contrário (erro). Este é um algoritmo de correção de erros. A

função de modulação será central na discussão futura. Por agora simplemente detecta erros.

Definimos

$$b_\mu := \frac{1}{\sqrt{N}}\xi_\mu.\mathbf{B}, \quad h_\mu(t) := \frac{1}{\sqrt{N}}\xi_\mu.\mathbf{J}(t) \tag{1.3}$$

Podemos reescalar $\mathbf{B}$ para que $\mathbf{B}.\mathbf{B} = N$.

Supomos que exista $\mathbf{B}$ e uma margem $\kappa > 0$ tal que para todo $\mu = 1 \dots P$

$$\xi_\mu.\mathbf{B}\,\sigma_\mu \geq \sqrt{N}\kappa > 0 \tag{1.4}$$

ou $b_\mu\sigma_\mu \geq \kappa > 0$.

Num dado instante t, temos $F = \sum f_\mu$ e $\mathbf{J}(t) = \sum f_\mu\xi_\mu\,\sigma_\mu$, respectivamente o número de passos efetivos de aprendizagem e o estado dos pesos do perceptron. Para este teorema F é a quantidade central. Queremos mostrar que permance finita, ou seja a partir de um certo ponto os valores de f_μ serão nulos e o todas as condições de estabilidade satisfeitas.

Multiplicamos as equações 1.5 por f_μ e as somamos obtendo

$$\sum f_\mu\xi_\mu.\mathbf{B}\,\sigma_\mu = \sqrt{N}\mathbf{J}.\mathbf{B} \geq F\kappa \tag{1.5}$$

Introduzimos $\rho = \frac{\mathbf{J}.\mathbf{B}}{|\mathbf{J}||\mathbf{B}|}$. E fácil mostrar que ρ esta entre -1 e 1, pois é o cosseno do angulo entre os dois vetores. Assim temos

$$(F\kappa)^2 \leq \rho^2 N|\mathbf{J}|^2 \tag{1.6}$$

e da dinâmica, quando houve um erro ($f_\mu = 1$)

$$\begin{aligned}
|\mathbf{J}(t+1)|^2 &= |\mathbf{J}(t) + \frac{f_\mu}{\sqrt{N}}\sigma_\mu\xi_\mu|^2 && (1.7)\\
&= |\mathbf{J}(t)|^2 + 2\frac{f_\mu}{\sqrt{N}}\mathbf{J}(t).\xi_\mu\sigma_\mu + \frac{1}{N}|\xi_\mu|^2 && (1.8)\\
&= |\mathbf{J}(t)|^2 + 2f_\mu h_\mu(t)\sigma_\mu + 1 && (1.9)
\end{aligned}$$

Podemos usar a escala $|\xi_\mu|^2 = N$ porque a classificação não depende dos módulos. Com respeito ao exemplo μ, se tivesse sido corretamente classificado, então não haveria mudança. Portanto não o foi e $h_\mu\sigma_\mu < \kappa\sqrt{N}$.

$$\begin{aligned}
|\mathbf{J}(t+1)|^2 &\leq |\mathbf{J}(t)|^2 + 2\kappa + 1 && (1.10)\\
&\leq F(2\kappa+1) && (1.11)
\end{aligned}$$

pois a cada passo o amumento é menor que $2\kappa + 1$ e foram dados F passos efetivos de aprendizagem.

$$(F\kappa)^2 \leq N|\mathbf{J}(t+1)|^2 \leq NF(2\kappa+1) \tag{1.12}$$

$$F \leq N(2\kappa^{-1} + \kappa^{-2}) \tag{1.13}$$

1.2.2 Comentários

Fica provado que se existir uma margem κ e um vetor $\mathbf{B}$, será encontrado algum vetor que reproduz os rótulos σ_μ para o conjunto de vetores no conjunto de treinamento. E se não for satisfeita alguma dessas condições? Primeiro, se os vetores forem sorteados de acordo a alguma distribuiçao de probabilidades, por exemplo uniformemente, pode ser que alguns estejam arbitrariamente perto da borda de decisão, o hiperplano perpendicular ao vetor professor $\mathbf{B}$. Neste caso o valor de κ será arbitrariamente pequeno. Isso levará à possibilidade de aprender a regra mas não em tempo finito. O vetor $\mathbf{J}$ evoluirá, ficando cada vez mais perto de $\mathbf{B}$ e o erro de generalização cairá com o número de exemplos apresentados.

Se não existir um vetor $\mathbf{B}$, então temos algumas alternativas possíveis. Pode ser que não haja regra para a associação de rótulos aos vetores. Podemos encontrar, para número de exemplos P baixo em relação a N, direções $\mathbf{J}$ que classificam corretamente os exemplos. Isso ocorre quando os vetores não são maliciosamente escolhidos (forma geral) e $\alpha = \frac{P}{N}$ é menor que 1. Para $\alpha < 2$ ainda há uma probabilidade razoavel de poder separar os vetores nas duas classes. Acima de 2 fica cada vez mais difícil, para N's maiores, de separá-los linearmente. Este resultado foi primeiro obtido por Cover usando idéias de geometria combinatórial e posteriormente por Gardner usando o método de réplicas. No limite $N \to \infty$ com probabilidade 1 para $\alpha < 2$ e probabilidade 0 para $\alpha > 2$ ocorre separação linear por um hiperplano que pode ser encontrado com o algoritmo perceptron de Rosenblatt.

Em problemas de aplicações os vetores representam um conjunto de caracterísiticas de cada padrão a ser classificado. Se for possível aumentar o número de características N, mantendo P fixo e com isso levar o problema a casos linearmente separáveis, podemos usar o perceptron. Mas pode ser que essa estratégia seja impossível. Podemos então colocar vários perpceptrons cuja saida é combinada para formar vetores de entrada num perceptron de saida. Assim temos as chamadas máquinas de camada escondida. Pode se provar (ref) que se o número de camadas escondidas for arbitrário, qualquer classificação pode ser implementada.

1.3 Aprendizagem supervisionada online

Vários cenários de aprendizagem podem ser considerados com respeito 'a forma como os exemplos são usados. Podemos usar todos os exemplos ao mesmo tempo e estudaremos isso mais adiante, sob o nome de aprendizagem offline. Podemos apresentar em subgrupos de tamanho menor que o que tem o conjunto de aprendizado. Podemos apresentar cada vetor de forma sequencial, visitando ciclicamente o conjunto de treinamento. Analisaremos a seguir cenários de apresentação única de exemplos. O interesse neste tipo de situação é que pode servir de modelo para várias situações de aprendizagem, pode ser estudado de forma relativamente fácil, apresenta um desempenho muito bom e serve como exemplo

para várias propriedades bastante gerais de situações de aprendizado.

Estudaremos o caso de N ser suficientemente grande para que $N \to \infty$ seja uma aproximação razoável. Isso nos permitirá escrever a dinâmica de aprendizado através de um conjunto pequeno de equações diferenciais ordinárias, obtidas por primeira vez por Kinzel e Rujan no contexto de seleção de exemplos.

Após a obtenção destas equações, estudaremos qual seria a melhor escolha da função de modulação, para poder extrair mais informação do conjunto de treinamento.

1.3.1 Formulação de aprendizagem online com apresentação única de exemplos

Voltemos às N equações 1.2. Transformamos as N equações em duas equações que descrevem o que podemos considerar como candidatos a parâmetros de ordem adequados para a descrição da aprendizagem. Para isso olhamos para o módulo de $\mathbf{J}$ e para $\mathbf{B}.\mathbf{J}$ que devem servir para medir a semelhança entre o aluno, representado por $\mathbf{J}$ e a regra ou professor, dado por $\mathbf{B}$.

$$|\mathbf{J}(t+1)|^2 = |\mathbf{J}(t)|^2 + 2\frac{f_\mu}{\sqrt{N}}\sigma_\mu \mathbf{J}(t).\xi_\mu + \frac{f_\mu^2}{N}|\xi_\mu|^2 \tag{1.14}$$

$$\mathbf{B}.\mathbf{J}(t+1) = \mathbf{B}.\mathbf{J}(t) + \frac{f_\mu}{\sqrt{N}}\sigma_\mu \mathbf{B}.\xi_\mu \tag{1.15}$$

e definimos $Q = \frac{|\mathbf{J}|}{\sqrt{N}}$ e $\rho = \frac{\mathbf{B}.\mathbf{J}}{QN}$, e também para facilitar as equações abaixo $R = \rho Q = \frac{\mathbf{B}.\mathbf{J}}{N}$

$$\Delta Q^2 = Q(t+1)^2 - Q(t)^2 = \frac{1}{N}(2f_\mu \sigma_\mu h_\mu + f_\mu^2) \tag{1.16}$$

$$\Delta R = R(t+1) - R(t) = \frac{1}{N} f_\mu \sigma_\mu b_\mu \tag{1.17}$$

$$\Delta \rho = \rho(t+1) - \rho(t) = \frac{1}{N}\left(\sigma_\mu (b_\mu - \frac{\rho}{Q}h_\mu)\frac{f_\mu}{Q} - \frac{\rho}{2}(\frac{f_\mu}{Q})^2\right) \tag{1.18}$$

Tomamos agora o limite termodinâmico e usamos que os parâmetros de ordem Q^2 e R ou ρ adquirem seus valores macroscópicos devido à automediancia, pois são compostos de somas. Já as diferenças ΔQ^2 e ΔR se transforamam em derivadas. Definimos $\alpha = \mu/N$ e $d\alpha = 1/N$

$$\frac{\Delta Q^2}{1/N} = 2f_\mu \sigma_\mu h_\mu + f_\mu^2 \tag{1.19}$$

$$\frac{\Delta R}{1/N} = f_\mu \sigma_\mu b_\mu \tag{1.20}$$

$$\frac{\Delta \rho}{1/N} = \left(\sigma_\mu (b_\mu - \frac{\rho}{Q}h_\mu)\frac{f_\mu}{Q} - \frac{\rho}{2}(\frac{f_\mu}{Q})^2\right) \tag{1.21}$$

$$\frac{dQ^2}{d\alpha} = < \lim_{N\to\infty} \frac{\Delta Q^2}{1/N} > \quad (1.22)$$

$$\frac{dR}{d\alpha} = < \lim_{N\to\infty} \frac{\Delta R}{1/N} > \quad (1.23)$$

$$\frac{d\rho}{d\alpha} = < \lim_{N\to\infty} \frac{\Delta \rho}{1/N} > \quad (1.24)$$

$$\frac{dQ^2}{d\alpha} = < 2f\sigma h + f^2 > \quad (1.25)$$

$$\frac{dR}{d\alpha} = < f\sigma b > \quad (1.26)$$

$$\frac{d\rho}{d\alpha} = < \left(\sigma(b - \frac{\rho}{Q}h)\frac{f}{Q} - \frac{\rho}{2}(\frac{f}{Q})^2\right) > \quad (1.27)$$

onde a média é feita sobre o exemplo que foi usado no passo de aprendizagem.

exercício Mostre das equações acima que $Qd\rho/d\alpha + \rho dQ/d\alpha = dR/d\alpha$

Equações diferenciais ordinárias para descrever a dinâmica de aprendizagem foram introduzidas em uma caso específico $f = 1$ por Kinzel e Rujan e posteriormente generalizadas por OK e NC. Reents e Urbanczick mostram em (PRL 1998) que a dedução acima pode ser demostrada de forma rigorosa.

1.3.2 Distribuição dos campos no caso de exemplos uniformemente distribuidos

Precisamos calcular $P(b, h|\mathbf{B}, \mathbf{J})$ que escrevemos $P(b, h)$ por simplicidade:

$$\begin{aligned} P(b,h) &= \int d\mu(\xi)\delta(b - \frac{\xi.\mathbf{B}}{\sqrt{N}})\delta(h - \frac{\xi.\mathbf{J}}{\sqrt{N}}) \quad (1.28)\\ &= \int d\mu(\xi)\frac{d\hat{b}d\hat{h}}{(2\pi)^2}\exp -i\hat{b}(b - \frac{\xi.\mathbf{B}}{\sqrt{N}})\exp -i\hat{h}(h - \frac{\xi.\mathbf{J}}{\sqrt{N}})\\ &= \int \frac{d\hat{b}d\hat{h}}{(2\pi)^2}\exp -i(\hat{b}b + \hat{h}h)\int d\mu(\xi)\exp i(\frac{\xi.(\hat{b}\mathbf{B} + \hat{h}\mathbf{J})}{\sqrt{N}}) \end{aligned}$$

e pelo teorema do limite central segue que $P(b, h)$ é uma gaussiana correlacionada. Para o caso em que a distribuição de exemplos é uniforme ($P(\xi_i = 1) =$

$P(\xi_i = -1) = 1/2)$ temos:

$$\begin{aligned}
\int d\mu(\xi) \exp i(\frac{\xi.(\hat{b}\mathbf{B} + \hat{h}\mathbf{J})}{\sqrt{N}}) &= \prod_i^N \int d\xi P(\xi) \exp i(\frac{\xi_i(\hat{b}B_i + \hat{h}J_i)}{\sqrt{N}}) \\
&= \prod_i^N \cos(\frac{(\hat{b}B_i + \hat{h}J_i)}{\sqrt{N}}) \\
&= \prod_i^N \exp -\frac{1}{2}(\frac{(\hat{b}B_i + \hat{h}J_i)}{\sqrt{N}})^2 \\
&= \exp -\frac{1}{2}\frac{(\hat{b}^2\mathbf{B}^2 + 2\hat{b}\hat{h}\mathbf{B}.\mathbf{J} + \hat{h}^2\mathbf{J}^2)}{N} \\
&= \exp -\frac{1}{2}(\hat{b}^2 + 2\hat{b}\hat{h}\rho Q + \hat{h}^2Q^2) \qquad (1.29)
\end{aligned}$$

e só falta fazer a transformada de Fourier

$$\begin{aligned}
P(b,h) &= \int \frac{d\hat{b}d\hat{h}}{(2\pi)^2} \exp -i(\hat{b}b + \hat{h}h) \exp -\frac{1}{2}(\hat{b}^2 + 2\hat{b}\hat{h}\rho Q + \hat{h}^2Q^2) \\
&= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp -\frac{b^2 + 2\rho bhQ^{-1} + Q^{-2}h^2}{2(1-\rho^2)} \qquad (1.30) \\
P(h) &= \frac{1}{\sqrt{2\pi Q^2}} \exp -(\frac{h^2}{2Q^2}) \qquad (1.31) \\
P(b) &= \frac{1}{\sqrt{2\pi}} \exp(-\frac{b^2}{2}) \qquad (1.32) \\
P(h|b) &= \frac{1}{2\pi Q\sqrt{1-\rho^2}} \exp -\frac{(b - \rho hQ^{-1})^2}{2(1-\rho^2)} \qquad (1.33)
\end{aligned}$$

onde usamos $\mathbf{B}^2/N = 1$, $\mathbf{J}^2/N = Q^2$ e $\rho = \frac{\mathbf{B}.\mathbf{J}}{|\mathbf{B}||\mathbf{J}|}$

Assim temos $< b^2 >= 1, < h^2 >= Q^2, < bh >= \rho$.

1.3.3 Erro de generalização

O erro de generalização é uma ferramenta útil para medir o desempenho de um algoritmo de aprendizado e podemos calculá-lo, novamente para distribuições uniformes de exemplos, a partir de

$$\begin{aligned}
e_g &= \int d\mu(\xi)\Theta(-\mathbf{J}.\xi\,\mathbf{B}.\xi) \qquad (1.34) \\
&= \int d\mu(\xi)dbdh\delta(b - \frac{\xi.\mathbf{B}}{\sqrt{N}})\delta(h - \frac{\xi.\mathbf{J}}{\sqrt{N}})\Theta(-hb) \\
&= \int dbdhP(b,h)\Theta(-hb) \\
&= \frac{1}{\pi}\cos^{-1}\rho \qquad (1.35)
\end{aligned}$$

1.4 Algoritmos de aprendizado

Podemos estudar as equações diferenciais 1.27 para alguns casos especiais de algoritmos de aprendizagem. Os mais conhecidos são

- Algoritmo de Hebb, $f_H = 1$
- Algoritmo Perceptron de Rosenblatt, $f_P = \Theta(-bh)$
- Adaline
- AdaTron

Mais do que dar énfase aos resultados em particular que cada algortimo fornece, estudos deste tipo levam a questões sobre vantagens de um algoritmo sobre outro. Qual atinge o menor erro de generalização? Qual é mais eficiente? Qual é mais robusto ante mudanças das hipóteses usadas para construir este cenário de aprendizagem modelo ? Qual pode ser generalizado para problemas não linearmente separáveis? A seguir estudamos o problema de generalização ótima, feito em situação bastante artificial e simplificada mas que dá alguns resultados interessantes.

1.5 Algoritmo de generalização ótima

Para a dinâmica dada pela equação 1.2 queremos encontrar f de forma a ter maior ganho de informação por exemplo. A primeira pergunta que pode ser feita é sobre o espaço que f vive. Podemos de forma geral olhar para todas as variáveis do problema e definir dois grupos

- Conjunto das variáveis acessíveis ou visíveis $\mathcal{V}$
- Conjunto das variáveis não acessíveis ou escondidas $\mathcal{H}$

É óbvio que para definir uma situação minimamente realista precisamos nos restringir a funções de modulação que dependam somente das variáveis $\mathcal{V}$. As médias que aparecem na equação 1.27 podem ser consideradas sobre a distribuição $P(\mathcal{V}, \mathcal{H}) = P(\mathcal{V})P(\mathcal{H}|\mathcal{V})$ e portanto podemos escrever a equação para a dinâmica

$$\begin{aligned} \frac{d\rho}{d\alpha} &= < \left(\sigma(b - \frac{\rho}{Q}h)\frac{f}{Q} - \frac{\rho}{2}(\frac{f}{Q})^2 \right) >_{\mathcal{V},\mathcal{H}} \\ \frac{d\rho}{d\alpha} &= << \sigma(b - \frac{\rho}{Q}h) >_{\mathcal{H}|\mathcal{V}} \frac{f}{Q} - \frac{\rho}{2}(\frac{f}{Q})^2 >_{\mathcal{V}} \end{aligned} \tag{1.36}$$

A derivada $\frac{d\rho}{d\alpha}$ é um funcional da função de modulação f e portanto podemos tomar uma derivada funcional para calcular o máximo de aumento médio de ρ a cada passo da dinâmica. Dado que o erro de generalização depende unicamente

de ρ, estaremos maximizando a queda média de e_G. Obtemos que a máxima generalização on-line é obtido com a função de modulação

$$f_o =< Q\sigma(\frac{b}{\rho} - \frac{h}{Q}) >_{\mathcal{H}|\mathcal{V}} \tag{1.37}$$

A equação para ρ pode ser escrita

$$\frac{d\rho}{d\alpha} = \frac{\rho}{Q^2} < f_o f - \frac{1}{2} f^2 >_{\mathcal{V}} \tag{1.38}$$

e para a escolha de $f = f_0$

$$\frac{dQ^2}{d\alpha} = < 2f_o < \sigma h >_{\mathcal{H}|\mathcal{V}} + f_o^2 >_{\mathcal{V}} \tag{1.39}$$

$$\frac{dR}{d\alpha} = < f_o < \sigma b >_{\mathcal{H}|\mathcal{V}} >_{\mathcal{V}} \tag{1.40}$$

$$\frac{d\rho}{d\alpha} = \frac{\rho}{2Q^2} < f_o^2 >_{\mathcal{V}} \tag{1.41}$$

$$\frac{1}{Q}\frac{dQ}{d\alpha} = \frac{1}{2} < (\frac{< b >_{\mathcal{H}|\mathcal{V}}}{\rho})^2 - (\frac{< h >_{\mathcal{H}|\mathcal{V}}}{Q})^2 >_{\mathcal{V}} \tag{1.42}$$

$$\frac{1}{\rho}\frac{d\rho}{d\alpha} = \frac{1}{2} < (\frac{< b >_{\mathcal{H}|\mathcal{V}}}{\rho})^2 + (\frac{< h >_{\mathcal{H}|\mathcal{V}}}{Q})^2 - 2(\frac{< b >_{\mathcal{H}|\mathcal{V}}}{\rho})(\frac{< h >_{\mathcal{H}|\mathcal{V}}}{Q}) >_{\mathcal{V}}$$

Mostraremos que sob certas circunstâncias

$$\frac{1}{Q}\frac{dQ}{d\alpha} - \frac{1}{\rho}\frac{d\rho}{d\alpha} = - < (\frac{< h >_{\mathcal{H}|\mathcal{V}}}{Q})^2 >_{\mathcal{V}} + < (\frac{< b >_{\mathcal{H}|\mathcal{V}}}{\rho})(\frac{< h >_{\mathcal{H}|\mathcal{V}}}{Q}) >_{\mathcal{V}}$$

é zero, portanto $\frac{1}{Q}\frac{dQ}{d\alpha} = \frac{1}{\rho}\frac{d\rho}{d\alpha}$, o que pode permitir evaluar ρ, medindo Q, que não precisa de informação sobre a regra para ser medido.

Defina

$$\Gamma = \frac{\sqrt{1-\rho^2}}{\rho} = \tan(\pi e_G) \tag{1.43}$$

$$Dx = \frac{dx}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{1.44}$$

$$H(y) = \int_y^\infty Dx = \frac{1}{2}\,\mathrm{erfc}(\frac{x}{\sqrt{2}}) \tag{1.45}$$

A equação para ρ pode ser escrita

$$\frac{d\rho}{d\alpha} = \frac{1-\rho^2}{2\pi\rho} \int_{-\infty}^{\infty} Dx \frac{\exp(-x^2/\Gamma^2)}{H(x/\Gamma)} \tag{1.46}$$

onde usamos o resultado:

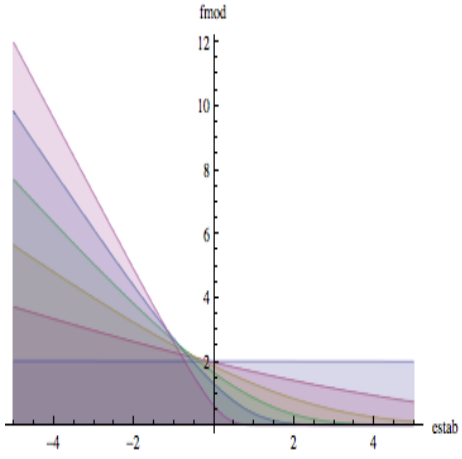

Figura 1.1: Função modulação de perceptron como função de $\sigma h/Q$ para diferentes valores de ρ (0.. Regra linearmente separável e estática, distribuição uniforme de exemplos;

Exercício Mostre que a funçao de modulação no caso do perceptron, é dada por

$$f_o(h,\rho,\sigma,Q) = \frac{Q\Gamma}{\sqrt{2\pi}} \frac{e^{-\frac{1}{2}(\frac{h}{Q\Gamma})^2}}{H(-\frac{\sigma h}{Q\Gamma})} \tag{1.47}$$

Exercício Mostre que assintoticamente $e_G \approx C\alpha^{-1}$. Mostre ainda que a constante $C \approx 0.88$

1.5.1 Análise da função de modulação

A forma da função modulação é interessante. Ela determina a intensidade com que o termo puramente hebbiano muda o vetor de pesos. Se σ e h são visíveis, então

$$f_o = \frac{Q}{\rho}\sigma\left(< b >_{\mathcal{H}|\mathcal{V}} - \frac{\rho}{Q}h\right) \tag{1.48}$$

e f_o funciona quase como se estivesse tentando corrigir erros. Calcula $< b >_{\mathcal{H}|\mathcal{V}}$, se este valor fosse comparado com h/Q, então seria um puro corretor de erros. Mas não é exatamente isso que é feito. A comparação é com $\rho h/Q$. Ou seja, leva em conta que h/Q e b não são perfeitamente correlacionados. Isso faz com que para os estágios iniciais de aprendizagem, quando ρ é pequeno, o valor de h não seja muito importante.

Assim a função de modulação (fig **??**) é para valores pequenos de ρ é uma constante. Aprendizagem leva a aumento de ρ e a modulação da mudança dos pesos diminui para exemplos corretamente classificados mas aumenta para padrões em que a previsão se mostra errada.

1.5.2 Seleção de Exemplos

Se os exemplos não forem escolhidos de forma uniforme, a função de modulação ainda sera a mesma, pois a distribuição relevante é $P(b|h)$, mas a escolha de exemplos nos permite aumentar em muito o desempenho do algoritmo de aprendizagem. A equação da dinâmica, para uma distribuição qualquer de exemplos $P(h)$ pode ser escrita como

$$\frac{d\rho}{d\alpha} = \frac{1-\rho^2}{2\pi\rho}\int_{-\infty}^{\infty} dh P(h)\exp(-h^2/\Gamma^2)\left[\frac{1}{H(x/\Gamma)} + \frac{1}{H(x/\Gamma)}\right] \tag{1.49}$$

Note que o fator que multiplica $P(h)$ é positivo, simétrico ante a troca $h \to -h$ e ainda tem um máximo para $h = 0$. Podemos escolher a distribuição de exemplos que maximiza $\frac{d\rho}{d\alpha}$:

$$P(h) = \delta(h) \tag{1.50}$$

Isto significa que se os exemplos com a menor margem possível forem escolhidos pela rede aluno a taxa de decaimento do erro de generalizção será máxima. Kinzel e Ruján sugeriram de froma heuristica que esta seria a melhor forma de selecionar os exemplos. A prova acima, olhando para a a equação da dinâmica, foi dada por KiCa . A equação diferencial fica simples, pois a delta faz a integral e obtemos

$$\frac{d\rho}{d\alpha} = 2\frac{1-\rho^2}{\pi\rho} \tag{1.51}$$

com a condição incial $\rho = 0$, temos

$$\rho = \sqrt{1 - e^{-2\alpha/\pi}} \tag{1.52}$$

Aqui vemos a vantagem de usar o algoritmo variacional. A seleção de exemplos leva a um algoritmo que converge exponencialmente rápido, mas isso não ocorre para o algorimo hebbiano estudado por Kinzel e Ruján, que continua decaindo com $\alpha^{-1/2}$ mesmo selecionando os exemplos. A única melhora ocorre no prefator constante, que cai à metade.

1.6 Aprendizagem na presença de ruído

Há várias formas de generalizar o modelo simples de aprendizado descrito acima. Podemos ter uma regra que muda no tempo, de froma deterministica ou aleatória, ou a comunicação entre a rede que aprende e aa fonte dos rótulos σ_B pode ser imperfeita. Nesta seção consideraremos dois tipos de modelos para o ruído, aditivo e multiplicativo. Denotamos o rótulo que chega ao aluno τ e abaixo descreveremos como deoende de σ. Além do erro de generalização

$$e_G =< \Theta(-h\sigma >_\xi \tag{1.53}$$

devemos considerar o erro de predição

$$e_P =< \Theta(-h\tau >_{\xi.\tau}> \tag{1.54}$$

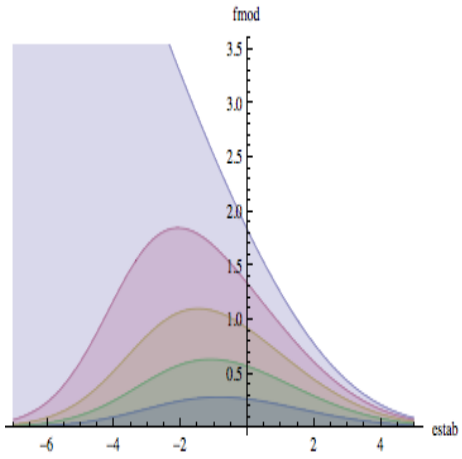

Figura 1.2: Função modulação de perceptron como função de $\sigma h/Q$ e ruido multiplicativo ϵ ($\rho = 0,4$. Regra linearmente separável e estática, distribuição uniforme de exemplos;

mostraremos, dentro do contexto de alguns modelos, que embora o erro de predição não possa ser reduzido a zero pelo processo de aprendizado, o erro de generalização ainda pode convergir a zero na presença de ruído.

Nestas condições o resultado da seção anterior muda. A média nas equações da dinâmica devera ser feita sobre o rótulo ruidoso, mas a idéia de otimizar ainda fornece resultados interessantes. A equação $f_o =< Q\sigma(\frac{b}{\rho} - \frac{h}{Q}) >_{\mathcal{H}|\mathcal{V}}$ ainda vale, mas mudam o conjunto de variáveis escondidas $\mathcal{H}$, que passa a incluir σ e o de variáveis acessíveis $\mathcal{V}$, que passa a ter τ

1.6.1 Ruído Multiplicativo

Com probabilidade $\eta \leq 1/2$ o rótulo σ é invertido: $P(\tau = \sigma) = 1 - \eta$, e $P(\tau = -\sigma) = \eta$,.

Os erros de predição e generalização estão relacionados, pois para qualquer função

$$< V(h,b,\tau) >_{P(h,b,\tau)} = \eta < V(h,b,-sign(b) >_{P(h,b)} + (1-\eta) < V(h,b,sign(b) >_{P(h,b)}$$

e portanto $e_P = \eta(1 - e_g) + (1 - \eta)e_G = \eta + (1 - 2\eta)e_G$, que sempre fica acima do nível η de ruído η

Hebb com ruído

Primeiro olhamos para aprendizagem hebbiano. As equações da dinâmica são

$$\frac{dQ^2}{d\alpha} = \sqrt{\frac{2}{\pi}}(1-2\eta) \tag{1.55}$$

$$\frac{d\rho}{d\alpha} = 2\sqrt{\frac{2}{\pi}}(1-2\eta)\rho + 1 \tag{1.56}$$

A dinâmica de Hebb é tão simples que podemos encontrar uma solução analítica.

Exercício Obtenha as equações acima (1.56) e mostre que

$$e_G(\alpha) = \frac{1}{\pi}\arccos\left[\left(1+\frac{\pi}{2(1-2\eta)^2\alpha}\right)^{-\frac{1}{2}}\right] \tag{1.57}$$

Usando que para $\rho \to 1$ o erro vai a zero e podemos olhar o comportamento assintótico, $\cos\pi e_G = 1-(\pi e_G)^2/2\rho$

$$e_G(\alpha) = \frac{C(\eta)}{\sqrt{\alpha}}, \tag{1.58}$$

com $C(\eta) = 1/\sqrt{2\pi(1-2\eta)}$. Para $\eta < 1/2$ a regra é aprendida lentamente mas totalmente. A medida que o nível de ruído aumenta o prefator $C(\eta)$, que não depende de α aumenta e diverge em $\eta = 1/2$, refletindo o fato que não há informação nenhuma nos exemplos que permita inferir a regra.

Algoritmo ótimo com ruído multiplicativo

Primeiro precisamos

$$P(b|h,\tau) = \frac{P(b,\tau|h)}{P(\tau|h)} = \frac{P(\tau|b,h)P(b|h)}{P(\tau|h)} \tag{1.59}$$

pois b não é acessível e τ e h o são. Assim, como τ, dado b, não depende de h, se reduz a

$$P(b|h,\tau) = \frac{P(\tau|b)P(b|h)}{P(\tau|h)}P(\tau|h) \tag{1.60}$$

Usando que $P(\tau|b)$ e $P(\tau|h)$ são as marginais de $P(\sigma,\tau|b)$ e $P(\sigma,\tau|h)$, podemos escrever

$$P(\tau|b) = \sum_\sigma P(\sigma,\tau|b) = \sum_\sigma P(\tau|\sigma,b)P(\sigma|b) \tag{1.61}$$

$$= \sum_\sigma P(\tau|\sigma)P(\sigma|b) \tag{1.62}$$

$$P(\tau|h) = \sum_\sigma P(\sigma,\tau|h) = \sum_\sigma P(\tau|\sigma,h)P(\sigma|h) \tag{1.63}$$

$$= \sum_\sigma P(\tau|\sigma)P(\sigma|h) \tag{1.64}$$

É fácil ver que $P(\sigma|b) = \Theta(-\sigma b)$ e $P(\sigma|h) = H(-\sigma h/\Gamma)$, o que leva a

$$P(\tau|b) = \eta + (1 - \frac{\eta}{2})\Theta(-\sigma b) \quad (1.65)$$

$$P(\tau|h) = \eta + (1 - \frac{\eta}{2})H(-\sigma h/Q\Gamma) \quad (1.66)$$

$$P(b|h,\tau) = P(b|h)\frac{(1-2\eta)\Theta(-\sigma b)}{\eta + (1-2\eta)H(-\sigma h/Q\Gamma)} \quad (1.67)$$

que vale para a istribuição uniforme de exemplos. Assim é possivel chegar à forma da função de modulção que aprece na figura 1.2:

$$f(h,\tau,Q,\rho,\eta) = \frac{Q\Gamma}{\sqrt{2\pi}}(1-2\eta)\frac{e^{-(\frac{h}{Q\Gamma})^2}}{\eta + (1-2\eta)H(-\sigma h/(Q\Gamma))} \quad (1.68)$$

1.6.2 Estimativa do ruido

1.7 Complexidade dos algoritmos e Ordenamento temporal