دانشکده‌ی مهندسی کامپیوتر

یادگیری ماشین بهار ۱۴۰۰

پاسخ امتحان میان‌ترم

مدرس: دکتر محمّدحسین رهبان

سوال ۱ رگرسیون لجستیک

آ)

$$p(y_i|x_i;w) = f(x_i;w)^{\frac{1+y_i}{2}}(1-f(x_i;w))^{\frac{1-y_i}{2}}$$

$$L(w) = \prod_{i=1}^{N} p(y_i|x_i;w) \Rightarrow \mathcal{L}(w) = \log(L(w)) = \sum_{i=1}^{N} \log p(y_i|x_i;w) \Rightarrow$$

$$\mathcal{L}(w) = \sum_{i=1}^{N} \log(f(x_i;w)^{\frac{1+y_i}{2}}(1-f(x_i;w))^{\frac{1-y_i}{2}}) = \frac{1}{2}\sum_{i=1}^{N}[(1+y_i)\log(f(x_i;w)) + (1-y_i)\log(1-f(x_i;w))]$$

$$f(x_i;w) = \frac{1}{1+e^{-w^\top x_i}} \Rightarrow 1 - f(x_i;w) = \frac{e^{-w^\top x_i}}{1+e^{-w^\top x_i}} = \frac{1}{1+e^{w^\top x_i}} \Rightarrow$$

$$\mathcal{L}(w) = \frac{1}{2}\sum_{i=1}^{N}[(1+y_i)\log(\frac{1}{1+e^{-w^\top x_i}}) + (1-y_i)\log(\frac{1}{1+e^{w^\top x_i}})] \Rightarrow$$

$$\mathcal{L}(w) = -\frac{1}{2}\sum_{i=1}^{N}[(1+y_i)\log(1+e^{-w^\top x_i}) + (1-y_i)\log(1+e^{w^\top x_i})]$$

ب)

$$f(x) = \log(1+e^x) \Rightarrow f'(x) = \frac{1}{1+e^x}\frac{de^x}{dx} = \frac{e^x}{1+e^x}$$

$$f''(x) = \frac{d}{dx}(\frac{e^x}{1+e^x}) = \frac{e^x}{1+e^x} - \frac{e^x \times e^x}{(1+e^x)^2} = \frac{e^x(1+e^x) - e^{2x}}{(1+e^x)^2} = \frac{e^x}{(1+e^x)^2} \geq 0$$

تابع دو بار مشتق پذیر بوده و مشتق دوم آن مثبت است در نتیجه محدب است.

پ)

$$h(x) = g(f(x)) \Rightarrow h'(x) = g'(f(x))f'(x) \Rightarrow h''(x) = g''(f(X))f'^2(x) + g'(f(x))f''(x)$$

در عبارت آخر به خاطر محدب بودن توابع $f(x), g(x)$ می‌دانیم که $g''(x), f''(x) > 0$ هستند. با توجه به اینکه $f'^2(x) \geq 0$ است اگر $g'(f(x)) > 0$ باشد در آن صورت $h''(x) > 0$ خواهد بود که طبق قسمت قبل به معنی محدب بودن آن است. پس شرط کافی برای محدب بودن $g(f(x))$ این است که $g'(x) > 0$ باشد یعنی تابع $g(x)$ باید صعودی باشد.

ت)

$$h(x) = \sum_{i=1}^{K} c_i f_i(x), \qquad f_i(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f_i(x_1) + (1-\alpha) f_i(x_2) \Rightarrow$$

$$h(\alpha x_1 + (1-\alpha)x_2) = \sum_{i=1}^{K} c_i f_i(\alpha x_1 + (1-\alpha)x_2) \leq \sum_{i=1}^{K} c_i(\alpha f_i(x_1) + (1-\alpha) f_i(x_2)) \Rightarrow$$

$$h(\alpha x_1 + (1-\alpha)x_2) \leq \alpha \sum_{i=1}^{K} c_i f_i(x_1) + (1-\alpha) \sum_{i=1}^{K} c_i f_i(x_2) = \alpha h(x_1) + (1-\alpha) h(x_2) \Rightarrow$$

$$h(\alpha x_1 + (1-\alpha)x_2) \leq \alpha h(x_1) + (1-\alpha) h(x_2)$$

پس ترکیب خطی تعدادی تابع محدب با ضرایب مثبت، محدب خواهد بود.

ث) برای اثبات مقعر بودن $\mathcal{L}(w)$ کافی است محدب بودن $\mathcal{M}(w) = -\mathcal{L}(w)$ را اثبات کنیم.

$$\begin{aligned}\mathcal{M}(w) &= \frac{1}{2} \sum_{i=1}^{N} [(1+y_i)\log(1+e^{-w^\top x_i}) + (1-y_i)\log(1+e^{w^\top x_i})] \\ &= \sum_{i=1}^{N} [a_i \log(1+e^{-w^\top x}) + b_i \log(1+e^{w^\top x})] = \sum_{i=1}^{N} [a_i \log(1+e^{-w^\top x})] + \sum_{i=1}^{N} [b_i \log(1+e^{w^\top x})]\end{aligned}$$

با توجه به این که برچسب‌ها $1+$ یا $1-$ هستند، ضرایب $a_i, b_i \geq 0$ خواهند بود. حال اگر نشان دهیم که که توابع $h_1(x) = \log(1+e^{w^\top x})$ و $h_2(x) = \log(1+e^{-w^\top x})$ محدب هستند، با توجه به نتیجه قسمت ت و مثبت بودن a_i, b_iها می‌توانیم نتیجه بگیریم که $\mathcal{M}(w)$ محدب است.

$$g(x) = \log(1+e^x), \quad f_1(x) = w^\top x \Rightarrow h_1(x) = g(f_1(x)) = \log(1+e^{w^\top x})$$

$$g(x) = \log(1+e^x), \quad f_2(x) = -w^\top x \Rightarrow h_2(x) = g(f_2(x)) = \log(1+e^{-w^\top x})$$

حال طبق تعریف توابع محدب:

$$f_1(x) = w^\top x \Rightarrow f_1(\alpha x_1 + (1-\alpha)x_2) = \alpha w^\top x_1 + (1-\alpha) w^\top x_2 = \alpha w^\top x_1 + (1-\alpha) w^\top x_2 \Rightarrow$$

$$f_1(\alpha x_1 + (1-\alpha)x_2) = \alpha f_1(x_1) + (1-\alpha) f_1(x_2)$$

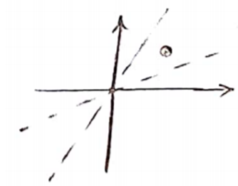
$$f_2(x) = -w^\top x \Rightarrow f_1(\alpha x_1 + (1-\alpha)x_2) = -\alpha w^\top x_1 - (1-\alpha) w^\top x_2 = -\alpha w^\top x_1 - (1-\alpha) w^\top x_2 \Rightarrow$$

$$f_2(\alpha x_1 + (1-\alpha)x_2) = \alpha f_2(x_1) + (1-\alpha) f_2(x_2)$$

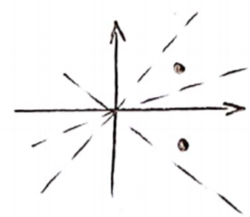
هر دو تابع در شرط محدب بودن صدق می‌کنند پس توابعی محدب هستند. از آنجایی که $g'(x) = \frac{e^x}{1+e^x} > 0$ است، یعنی تابعی صعودی است پس طبق نتیجه قسمت پ دو تابع $h_1(x), h_2(x)$ توابعی محدب هستند. پس تابع $\mathcal{M}(w)$ یک تابع محدب بوده و در نتیجه تابع $\mathcal{L}(w)$ یک تابع مقعر است.

سوال ۲ دسته‌بندی دودویی در فضای دوبعدی

آ) $m_H(1) = 2$

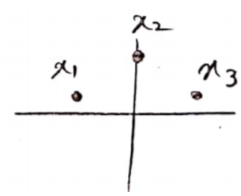


$m_H(2) = 4$

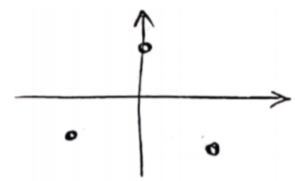


$m_H(3) = 6$

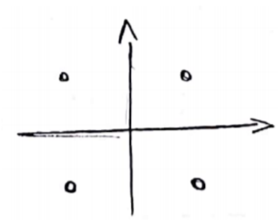
اگر داده‌ها در یک طرف خط باشند، حالتی که داده‌ها یکی در میان برچسب گذاری(+-+ یا -+-) شده باشند قابل تولید نیست.



اگر داده‌ها در دو طرف خط توزیع شده باشند، حالتی که داده‌ها هم برچسب(هر سه مثبت یا هر سه منفی) باشند قابل تولید نیست.



$m_H(4) = 8$ چهار نقطه را به شکل زیر میچینیم، اگر بخواهیم فقط یک نقطه را علامت متفاوت بدهیم چهار حالت قابل جداسازی نیستند، همچنین دو حالت تمام مثبت و تمام منفی هم ایجاد نمیشوند، همینطور دو حالتی که به صورت ضربدری علامت دهی شده باشند قابل جداسازی با این کلاس نیستند. بنابراین 8 حالت را نمیتوان ایجاد کرد.



ب) دو تابع را f و g می‌نامیم. هر کدام از f و g حداکثر یک عنصر به $M(x_1,...x_N)$ اضافه می‌کنند، بنابراین هر کدام حداکثر دو حالت اضافه می‌کنند:

$$\max_{f,g} m_M(N) = \begin{cases} 2 & N=1 \\ 4 & N=2 \\ 8 & N=3 \\ 10 & N=4 \end{cases}$$

پ) هر کدام از فرمول‌هایی که برای خطای تعمیم بیان شده را استفاده کنید به شرط آنکه درست جایگذاری کرده باشید نمره کامل را خواهید گرفت. بعضی از فرمول ها به شکل زیر هستند:

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N} = \frac{2m_M(N)}{e^N}$$

یا

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 4m_M(2N)e^{-\frac{1}{8}\epsilon^2 N} = \frac{4m_M(2N)}{e^{\frac{N}{16}}}$$

از طرف دیگر با توجه به بخش **ب** میدانیم $m_M(N) \leq N^{d_{vc}(M)} + 1 = N^3 + 1$ بنابراین:

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq \frac{4((2N)^3 + 1)}{e^{\frac{N}{16}}}$$

حل با استفاده از باند VC نیز قابل قبول است.

ت) $E_{in}(h_1) = \sum_{n=1}^{N} |h_1 - y_n|$ را کمینه می‌کنیم:

$$\frac{dE_{in}(h_1)}{dh_1} = \sum_{n=1}^{N} (h_1 - y_n) = 0$$

بنابراین اگر تعداد جمله های مثبت و منفی با هم برابر باشد، مقدار مشتق برابر صفر خواهد شد، پس:

$$h = median\{y_1, ..., y_N\} = h_{med}$$

و همچنین چون تخمینگر میانه داده‌ها است، داده‌های خارج از محدوده در مقدار آن تاثیر ندارند و بنابراین در صورت نویزی شدن y_N ، h_{med} تغییری نمی‌کند.

سوال ۳ تابع هزینه

آ) ۱. در موارد بسیاری این کار تاثیر عکس می‌گذارد در خروجی. اگر خط جداکننده را در نظر بگیریم، نقاطی که به درستی دسته‌بندی شده‌اند با فاصله گرفتن از این خط جریمه تولید می‌کنند و باعث می‌شوند خط جداکننده به سمت آن‌ها متمایل شود. این در حالیست که با

دسته‌بندی درست این نقاط، نباید جریمه‌ای حاصل شود. تصویر زیر از کتاب *Bishop* این موضوع را به خوبی به تصویر کشیده است.

186 4. LINEAR MODELS FOR CLASSIFICATION

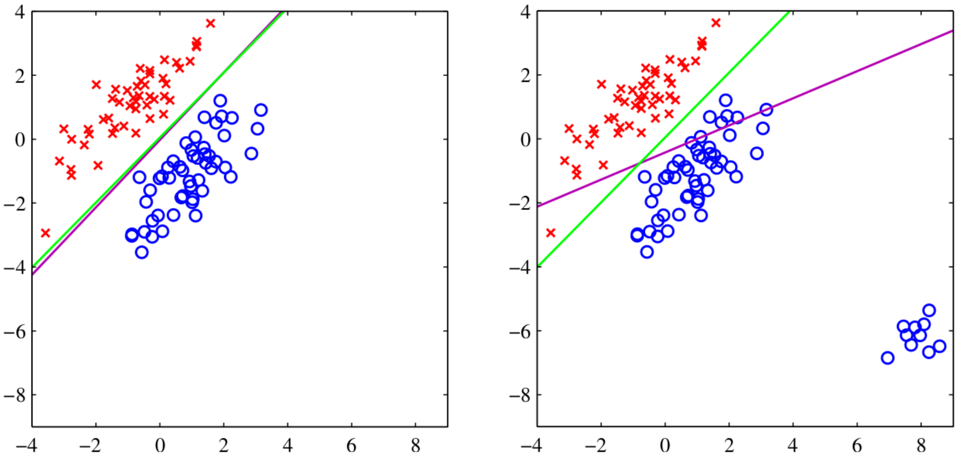


Figure 4.4 The left plot shows data from two classes, denoted by red crosses and blue circles, together with the decision boundary found by least squares (magenta curve) and also by the logistic regression model (green curve), which is discussed later in Section 4.3.2. The right-hand plot shows the corresponding results obtained when extra data points are added at the bottom left of the diagram, showing that least squares is highly sensitive to outliers, unlike logistic regression.

البته در اینجا مدل رگرسیون خطی را در نظر گرفتیم. مدل‌های دیگر نیز می‌توانست استفاده شود و متناسب با اشکالات مطرح‌شده‌ی آن‌ها نمره داده شده است.

۲. تابع هزینه‌ی پرسپترون، داده‌ها را درست دسته‌بندی می‌کند اما بین تمام خطوطی که خروجی می‌دهد تفاوتی قائل نمی‌شود. اما *Logistic Regression* داده‌ها را بسته به فاصله به خط جریمه می‌کند و مرز مناسب‌تر (و نه *Random*) را گزارش می‌کند.

ب) ۱. با توجه به محدب بودن تابع کافی‌ست، جایی که مشتق آن صفر می‌شود را بیابیم. داریم:

$$L_{MSE}(\theta) = \frac{1}{n}\sum_i (x_i - \theta)^2 \implies \frac{d}{d\theta}L_{MSE}(\theta) = \frac{2}{n}\sum_i (x_i - \theta) = 0 \implies \theta = \frac{\sum_i x_i}{n}$$

به عبارت دیگر، بهینه در میانگین اعداد رخ می‌دهد. حال تابع MAE را کمینه می‌کنیم. همچنین منظور از $sgn(x)$ تابع علامت است که برای اعداد مثبت برابر با یک و برای اعداد منفی برابر با منفی یک است (برای صفر نیز صفر است). مشابه بالا داریم:

$$L_{MAE}(\theta) = \frac{1}{n}\sum_i |x_i - \theta| \implies \frac{d}{d\theta}L_{MAE}(\theta) = \frac{1}{n}\sum_i sgn(x_i - \theta) = 0 \implies \sum_i sgn(x_i - \theta) = 0$$

با توجه به تعریف تابع sgn معادله بالا نتیجه می‌دهد مقدار θ باید به گونه‌ای باشد که تعداد اعضای بیشتر از آن با کمتر از آن برابر باشد و این همان تعریف میانه است. پس جواب بهینه برای معادله‌ی بالا میانه‌ی اعداد است.

۲. تابع دوم برای زمانی که داده‌های پرت یا *outlier* داریم عملکرد بهتری خواهد داشت تا میانگین. البته میزان خطا همچنان بالا خواهد بود. اما از طرف دیگر، گرادیان آن در تمام نقاط یکسان خواهد بود. به این معنا که گرادیان برای زمانی که حتی *loss* کمی داریم نیز زیاد است و برابر با زمانیست که *loss* زیادی داریم و عملا از آن نمی‌توانیم استفاده کنیم. در حالیکه با استفاده از تابع اول می‌توانیم با استفاده از گرادیان آن به نقطه‌ی بهینه برسیم.

پ) زمانی که داده‌ها خطی‌جداپذیر باشند، $w^\star$ وجود دارد که به ازای آن داریم:

$$w^{\star\top} x_i > 0 \quad \forall x_i \in C_1$$

$$w^{\star\top} x_i < 0 \quad \forall x_i \in C_2$$

برقرار است و این بدان معناست که به ازای هر $\alpha > 0$ نیز $\alpha w^\star$ در آن صدق می‌کند. همچنین بیشینه کردن عبارت زیر معادل کمینه کردن خطا است:

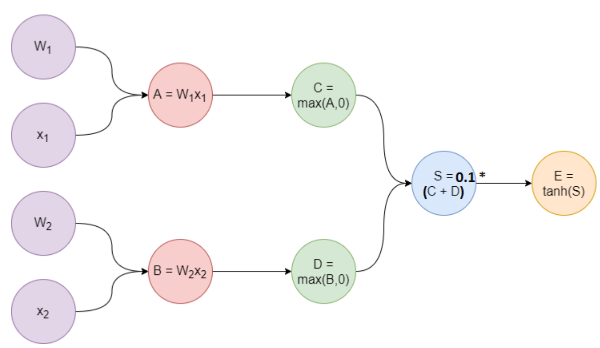
$$ln\, p(t|w) = \sum_i t_i ln\, y_i + (1 - t_i) ln(1 - y_i); \qquad y_i = \sigma(w^\top x_i)$$

حال از آنجا که $sigmoid$ تابع اکیداً صعودی‌ست هرچه w را بیشتر کنیم مقدار y_i برای کلاس مثبت مثبت‌تر و برای کلاس منفی منفی‌تر می‌شود. پس بیشینه مقدار عبارت به ازای $w = \infty w^\star$ رخ می‌دهد که این باعث $overfitting$ نیز می‌شود.

سوال ۴ شبکه عصبی

آ) اگر به داده های کمی برای آموزش دسترسی داریم بهتر است از مدل با پیچیدگی کمتر استفاده کنیم چرا که مدل با پیچیدگی بالا دچار *overfitting* به *deterministic noise* می شود. اگر به داده های زیادی برای آموزش دسترسی داریم می توانیم از مدل با پیچیدگی بالا استفاده کنیم.

ب)



$$A = \begin{bmatrix} -1 \\ 5 \end{bmatrix} \qquad B = \begin{bmatrix} 10 \\ -9 \end{bmatrix} \qquad C = \begin{bmatrix} 0 \\ 5 \end{bmatrix} \qquad D = \begin{bmatrix} 10 \\ 0 \end{bmatrix} \qquad S = \begin{bmatrix} 10 \\ 0.5 \end{bmatrix} \qquad E = \begin{bmatrix} 1 \\ 0.46... \end{bmatrix}$$

سپس برای مشتق داریم:

$$\frac{\partial E}{\partial E} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \frac{\partial E}{\partial S} = 1 - E^2 = \begin{bmatrix} 0 \\ 0.78... \end{bmatrix},$$

$$\frac{\partial E}{\partial C} = \frac{\partial E}{\partial S} \times \frac{\partial S}{\partial C} = 0.1 \times \frac{\partial E}{\partial S} = 0.1 \times \frac{\partial E}{\partial D} = \begin{bmatrix} 0 \\ 0.078... \end{bmatrix}$$

برای مشتق تابع max، مشتق تنها از نقاطی به عقب منتقل می شود که ورودی در آن نقاط از صفر بزرگتر بوده:

$$\frac{\partial E}{\partial A} = \frac{\partial E}{\partial C} \times \frac{\partial C}{\partial A} = \begin{bmatrix} 0 \\ 0.078... \end{bmatrix}, \quad \frac{\partial E}{\partial B} = \frac{\partial E}{\partial D} \times \frac{\partial D}{\partial B} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$\frac{\partial E}{\partial W_1} = \frac{\partial E}{\partial A} \times \frac{\partial A}{\partial W_1} = \frac{\partial E}{\partial A} \times x_1^\top = \begin{bmatrix} 0 \\ 0.078... \end{bmatrix} \times \begin{bmatrix} 1 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0.078... & 0.15... \end{bmatrix}$$

$$\frac{\partial E}{\partial x_1} = \frac{\partial E}{\partial A} \times \frac{\partial A}{\partial x_1} = W_1 \times \frac{\partial E}{\partial A} = \begin{bmatrix} -3 & 1 \\ 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0 \\ 0.078... \end{bmatrix} = \begin{bmatrix} 0.078... \\ 0.15... \end{bmatrix}$$

$$\frac{\partial E}{\partial W_2} = \frac{\partial E}{\partial B} \times \frac{\partial B}{\partial W_2} = \frac{\partial E}{\partial B} \times x_2^\top = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \times \begin{bmatrix} -2 & 3 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\frac{\partial E}{\partial x_2} = \frac{\partial E}{\partial B} \times \frac{\partial B}{\partial x_2} = W_2 \times \frac{\partial E}{\partial B} = \begin{bmatrix} -2 & 2 \\ 3 & -1 \end{bmatrix} \times \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

پ) حداکثر به دو لایه و مرتبه 2^n عدد راس پنهان نیاز داریم. در لایه اول تمامی $product$ های مورد نیاز را که از مرتبه 2^n هستند را می‌سازیم و برای خروجی همه آنها را با هم or می‌کنیم.

ث) کافی است وزن‌ها را قرار دهیم جوری که $h_i = f(x_i - x_{i+1})$ و همچنین $y = f(3 - h_1 - h_2 - h_3)$.

سوال ۵ منظم سازی

آ) با کاهش ضریب منظم‌ساز، خطای دادگان آموزش پایین می‌آید. همچنین با کاهش ضریب، اگر همچنان از حد خاصی بیشتر باشد خطا روی دادگان آزمون کاهش می‌یابد اما اگر این ضریب از حدی کمتر شود تاثیر منظم‌سازی کم شده و خطای آزمون افزایش می‌یابد.

ب) برای پیدا کردن منظم‌ساز بهینه باید این منظم‌ساز در جهت جواب مساله باشد و این مانند این است که مساله را حل کرده باشیم. در صورت استفاده از منظم‌ساز غلط با استفاده از $validation$ می‌توانیم پارامترش را تنظیم کنیم تا از حالت معمولی نتیجه بدتری نگیریم.

پ) نرم‌های ۱ بردارها به ترتیب 7 ، 7 و 5 اند و نرم‌های ۲ شان نیز به ترتیب $\sqrt{17}$ ، 5 و 5 هستند. بردارهای اول و دوم نرم ۱ یکسانی دارند، یعنی جمع مقادیرشان با هم برابر است. اما به دلیل فشرده و نزدیک بودن مقادیر بردار اول به یکدیگر، نرم ۲ی این بردار کمتر است. یعنی هرچه مقادیر از هم فاصله بگیرند و تعدادیشان بزرگ شوند، جمع توان دوم آنها که همان نرم ۲ است افزایش می‌یابد.

با مقایسه‌ی بردارهای دوم و سوم هم می‌توان دید که نرم ۲ی یکسان دارند. اما چون مقادیر بردار دوم به هم نزدیک‌اند نرم ۱ اش بیشتر است. بنابراین دور شدن قدر مطلق مقادیر از هم نرم ۲ را به صورت نسبی از نرم ۱ بیشتر می‌کند و این دور شدن، صفر شدن برخی از مقادیر را می‌تواند به همراه داشته باشد. در نتیجه با استفاده از منظم‌سازی که نرم ۱ را کمینه می‌کند ترجیح را بیشتر روی صفر کردن مقادیر می‌گذاریم. (توضیح ناقص‌تر با تکیه بر تعداد ۰ ها در مقایسه سه بردار نیز برای این سوال کافیست)

ت) طبق قضیه بیز می‌دانیم: $p(w|x,t,\alpha,\beta) \propto p(t|x,w,\beta)p(w|\alpha)$.

حال با استفاده از درست‌نمایی بیشینه برای بدست آوردن w می‌توانیم بنویسیم:

$$
\begin{aligned}
w^* &= argmax\ ln(p(w|x,t,\alpha,\beta)) = argmax\ ln(p(t|x,w,\beta)) + ln(p(w|\alpha)) \\
&= argmax\ ln(\prod \mathcal{N}(t_i|f(x_i,w),\beta^{-1})) + ln((\frac{\alpha}{2\pi})^{\frac{m+1}{2}} e^{-\frac{\alpha}{2}w^T w}) \\
&= argmax\ \frac{N}{2}ln(\beta) - \frac{N}{2}ln(2\pi) - \frac{\beta}{2}(\sum[t_i - f(x_i,w)]^2) + \frac{m+1}{2}ln(\frac{\alpha}{2\pi}) - \frac{a}{2}w^T w
\end{aligned}
$$

که N برابر با تعداد نمونه‌ها است. حال با حذف بخش‌های ثابت، بیشینه کردن عبارت بالا معادل با کمینه کردن عبارت زیر است که معادل با کمینه کردن خطای میانگین مربعات به علاوه‌ی منظم‌ساز L_2 می‌باشد.

$$
w^* = argmin\ \frac{\beta}{2}(\sum[t_i - f(x_i,w)]^2) + \frac{\alpha}{2}w^T w
$$

پیروز باشید