400108547 تمرین دوم یادگیری ماشین پرهام رضایی

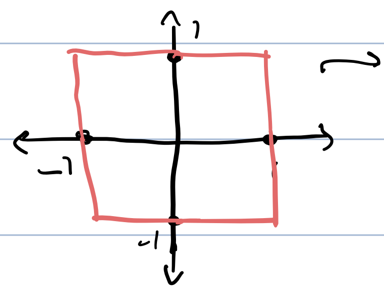
سوال یک

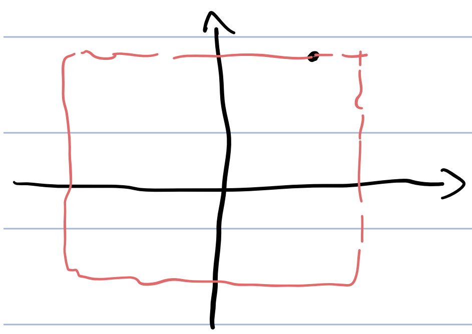
الف)

برای مثال بازه نقاط $[1,-1] \times [1,-1]$ را در نظر بگیرید.

مرز ناحیه نرم ها را رسم می کنیم.

برای (ابزار) (نرم) ← $l_\infty = 1$ چون که $|z| \le 1$





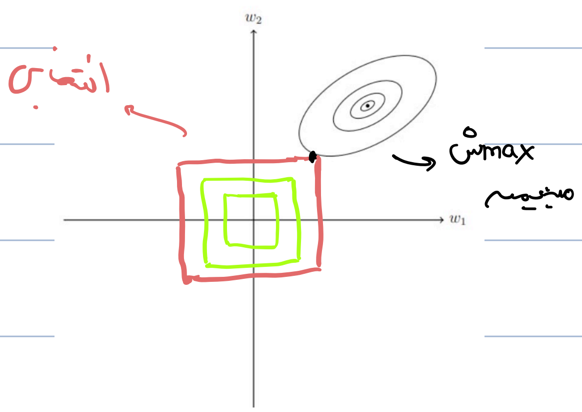
برای مجموعه این از وزن ها اگر یک نقطه را بگیرید.

حال مربع شامل این را (عرض و طول = l_∞ می شه)

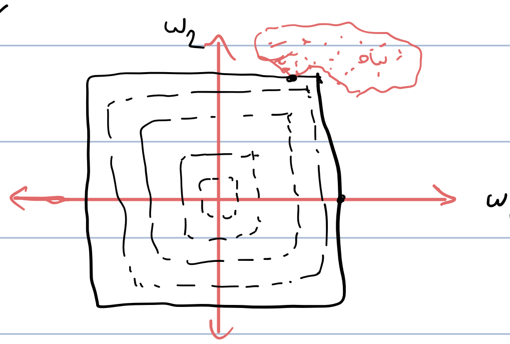
در نظر بگیریم سعی در کوچک کردن این مربع است.

پس نقطه انتخابی اگر از مرکز و مربع شروع کنیم و مربع را بزرگ کنیم نخستین

نقطه برخورد است ← که کوچکترین مربع برای این نقطه است.



شکل سوال:



l_∞ مینیمم ~

ب)

با استفاده از این نرم یعنی ∞ $\lambda \|w\|$ روی ماکسیمم اعضای w هزینه می‌گذاریم
که بهینه سبب ماکسیمم وزن‌ها خیلی بزرگ نشود.

معادل:

$$\min_{w} \; l(w,x,y) - \lambda \|w\|_\infty$$

$$\Longleftrightarrow \quad \min_{w} \; l(w,x,y)$$

$$\text{s.t} \quad \|w\|_\infty \leqslant C$$

یعنی روی ماکس وزن‌ها محدودیت قرار می‌دهیم.

در شرایطی که بخواهیم وزن‌ها محدودیت اندازه داشته باشند و از یک لیمیت خاص کمتر باشند
کاربرد دارند.

الف) ابتدا محدودیت ، بدون همخطی بیش پس احتمالاً (مسئله خود دچار جواب‌ها)

بایاس زیادی دارد. چه برسد با افزایش s که رگولاریزیشن روی وزن‌ها کمتر و

کمتر می‌شود پس بایاس کاهش می‌یابد اما ممکن است دچار اورفیت شویم

که در این شکل که فضای ترین هر نقطه مهم نیست.

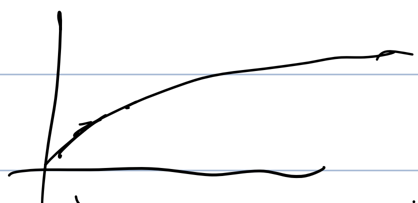
تا جایی که β_iها به بهینه خودشان برسند پس از آن پس s

اهمیتی ندارد و ثابت می‌شوند.

البته اگر متغیر y=0 باشه این‌ها رو صفره شاید برسه و کلاً در حالت کل کاهشیه.

پس 4 کاهش می‌یابد.

ب) با افزایش سقف محدودیت β_i داریم فضای فرضیه‌مان را گسترش می‌دهیم

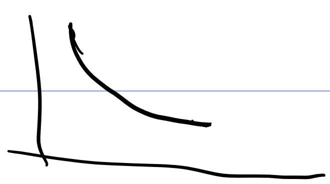


صفر چون همه صفرند اول

مدل پس واریانس افزایش می‌یابد.

پس 3

ج) ابتدا g با f بهینه چون H محدوده فاصله دارد ولی با گسترش H



کاهش می‌یابد. (حتی اگر f و g کلاً فاصله داشته باشند با افزایش s بهتر فیت نزدیک‌تر به دیتا می‌شود ⇐ کاهش)

پس 4

سوال ۳

فکر کنم بخش الف و ب سوال جابجا شد ولی جفتش هست :-{|

Data Augmentation

یکی از کاربردهای روتین در ویژن است.
وقتی دیتاستمون محدودیت هایی داره مثلا از نظر تعدادی یا جهت و ... برای اینکه مدل لرن شده تحت تاثیر اون محدودیت نشه می توانیم اگمنت شده های عکس ها رو هم به دیتاست اضافه کنیم. برای مثال تبدیل های افاین مختلف (روتیشن های مختلف هر عکس برای مثال) روی عکس ها زده یا کراپ شده بخش هایی یا ... را به دیتاست می افزاییم.
شهود: می دانیم برای مثال در یک تسک کلسیفیکیشن مثل تشخیص سگ :/ توی عکس توقع داریم زاویه عکس تاثیری در سگ بودنه نداشته باشد بدین ترتیب این رو به عنوان یک inductive bias با افزودن روتیشن های مختلف عکس به دیتاست به مدل اضافه میکنیم و توقع داریم مدل روی زاویه و ... تاثیر نگیره که اورفیت بشه و بعد روی دیتای جدید که لزوما همون زاویه نباشه غلط نشه.
کلا هم با اضافه کردن این دیتاها موجب میشیم نسبت به اون نویز یا حالا پترن روباست تر رفتار کنه.
از مثال روش های توی ویژن میشه افزودن دیتای اگمنتش شده تغییرهای crop,flip,rotation,translation,brightness,contrast,color,saturan اشاره کرد.
کلا توی خیلی تسکا نیاز به چیزایی عین سلف سوپروایزد هستش مثل خیلی مدلا در کانترستیو لرنینگ. دیتامون اساسا لیبل نداره میایم اگمنت شده های یه عکس رو توی دیتا قرار میدیم و سعی میکنیم امبد شدشونو نزدیک کنیم. اینجا اصلا لیبل نداریم واسه همین نیاز به دیتای این طوری داریم.
کارهای دیگه ای هم میشه کرد مثلا بیایم ادورسریال اتک هایی به دیتا بزنیم خود این اتک خورده ها رو هم به دیتا اضافه کنیم برای ترین مدل.
یا اگه دیتا کم داریم از مدلای جنریتیو مثل گن و ... استفاده کنیم که دیتا بیشتر بشود.
البته لزوما همیشه دیتا اگمنتیشن موجب بهبود اورفیت نمیشه.
حال دیتا اگمنتیشن های دیگه ای رو بررسی میکنیم. یکی از روش ها توی ام ال oversampling و undersampling هستش.
یکی از روش ها رندوم سمپلینگه. حالا یا اورسمپلینگ میکنیم که میایم برای کلاس ماینورمون دیتاهای تکراری استفاده میکنیم که این احتمال اورفیت رو میتونه زیاد کنه به خاطر دیتاهای تکراری. ازونور میتونیم اندرسمپلینگ رندوم کنیم از کلاس میجرمون که این میتونه موجب از دست رفتن اینفورمیشن مهم بشه.
روش های دیگه ای هم داریم مثل smote . اینجا میایم توی کلاس ماینور یه دیتارو میگیریم. k همسایه نزدیکش رو هم میگیریم.(k تعیین میکنیم معمولا ۵ مثلا) حالا یکی ازینا رو رندوم انتخاب میکنیم و روی پاره خط بین این دو تا (کلا دارم توی فضای این فیچرا بحث میکنم) یه نقطه جدید میگیریم و به دیتا اضافه می کنیم. اما این متودهم اگه واریانس دیتای ماینورمون زیاد باشه و شباهتا با میجره زیاد مشکل زاعه. همچنین کلا احتمال تولید دیتای نویزی و مشکلات اورفیت رو همچنان داره.
روش های دیگه اپسمپل مثل adasyn که بیسش همون smote هستش و یا اندرسمپلینگ مثل tomek links هست که میاد اورلپ کلاسای میجرو کم میکنه با حذف دیتاهایی که تومک لینکن و ...

یکم ساده لوحانه نگاه کنیم داریم که

rademacher complexity: $\hat{\mathcal{R}}_N(\mathcal{H}) = \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}}\left|\frac{1}{N}\sum_{i=1}^{N}\sigma_i h(x_i)\right|\right]$

$\mathcal{R}_N(\mathcal{H}) = \mathbb{E}\left[\hat{\mathcal{R}}_N(\mathcal{H})\right]$

حالا ارور جنرالیزیشنو داریم که :

$$\hat{\epsilon}_N(h) \le \mathcal{R}_N(\mathcal{H}) + \mathcal{O}\left(\sqrt{\frac{\ln {}^1\!/\!_\delta}{N}}\right)$$

با احتمال $1-\delta$

خب با این نگاه با افزایش N که تعداد سمپلامونه داره ارور جنرالیزیشنمون رو کم میکنه و خب یعنی داره از اورفیت جلوگیری میکنه ولی مشکلی که هست حداقل متودهای که ظاهر بیرونی رو حفظ میکنند برا ادم یعنی مثلا توی کلسیفیکیشن کلاسمون همونه استفاده زیاد از پرایار نالجمون درمورد دامنه دیتامون و ویژگی های تصویر مثلا داره خب iid نیستند و نمیشه به چشم دیتای جدید صاف بهشون نگاه کرد که خب نگاه تئوری لرنینگ طور بالامون اساسا همونطور که گفتم ساده لوحانه هستش. بررسی دقیق ریاضیش از توان من خارجه هعب.
یه پیپرم دیدم بررسی کرده بود اساسا اگه میتونید رگولاریزیشن استفاده کنید و قصد کاهش اورفیته تاثیر خیلی بهتر و قابل کنترل تر (از لحاظ تئوری) از دیتا اگمنتیشنه.

PCA and Dimension Reduction

پی سی ای و روشای کاهش ابعاد لزوما موجب کاهش اورفیت نمیشن. درسته بعضا میتونه کاهش بده ولی افزایشم ممکنه بده.
وقتی یه مدل فیت میکنیم داریم کوواریانس بین فیچرها و اون تارگت رو می یابیم و توی هدفمون کواریانس خود فیچرها با هم نیستش و دنبال تاثیر تغییرات فیچر روی تارگتیم.
حالا pca عملا به کواریانس با اون تارگته کاری نداره داره میاد واریانس فیچرارو بررسی میکنه و ازون حالا ایگن وکتورا رو میگیریم و ... . جهت بیشترین واریانس فضای فیچرمون لزوما دلیلی نداره ربط به کواریانس بین فیچرا و تارگت داشته باشه.
ازینکه pca بزنیم فیچرهای مهممون از دست بره تاثیرش.
یه مثال :

H = height of cloud-base
Ts = surface temperature
Td = surface drawpoint

حالا داریم که چون کواریانس دو تا فیچرا مثبته

$var(Ts+Td)=var(Ts)+var(Td)+2Cov(Ts,Td) > var(Ts) + var(Td)-2Cov(Ts,Td) > Var(Ts-Td)$

ولی H مستقیما به رطوبت ربط داره که از Ts-Td میاد. ولی اگه مثلا اینجا ما واریانس بیشترو نگه داریم دیگه Ts+Td به ما نتیجه خوبی توی پیشبینی H نمیده.

نکته: کلا این که واریانس کما نویز نیستند و اهمیت دارند واسه همین مدلای دیمنشن ریداکشنمون علی الخصوص pca لزوما اینجا خوب عمل نمیکنن و میتونن موجب اورفیت بشن توی این صفحه هم بررسی شده:
https://stats.stackexchange.com/questions/87198/low-variance-components-in-pca-are-they-really-just-noise-is-there-any-way-to/87231#87231

بررسی تئوری: به هر حال تعداد دیمنشنا داره کم میشه واسه همین کامپلکسیتی مدل کم شده این داره واریانس مدل رو کم میکنه که باعث کاهش اورفیت میتونه بشه ولی نمیشه تاثیرش روی بایاس هم صرف نظر کرد ممکنه کلا اینفورمیشن خوبمون رو از دست بدیم. پس میتونه باعث کاهش اورفیت بشه ولی لزوما نمیشه ممکنه زیادش کنه.
همچنین PCA و کلا متودای کاهش بعد میتونه موجب حذف نویز هم مقداری بشه که باعث کاهش اورفیت میشه.
ولی ازونور میتونه اینفورمیشن مهم از دست بده یا ریلیشن بین فیچرای مختلف رو مقداری بهم بریزه که درنتیجه اورفیت زیاد شه.
اینها برای کلا روشای کاهش بعد صدق میکنه چه خطی عین pca چه عین t-sne و ...

$$-l = \sum_{i=1}^{n} \log P(y_i \mid x_i, w_0, w_1, w_2) - C w_j^2$$

$$\Rightarrow l = -\sum_{i=1}^{n} \log P(y_i \mid x_i, w_0, w_1, w_2) + C w_j^2 \qquad \text{minimize}$$

* برای حل می‌توانید مستقیم به بروید

$$\rightarrow l = -\sum_{i=1}^{n} \Big(y_i \log P(1 \mid x^i, w_0, w_1, w_2) + (1-y_i) P(0 \mid x^i, w_0, w_1, w_2) \Big) + C w_j^2$$

$$= \sum_{i=1}^{n} \Big(-y^i \log \overbrace{\sigma(w^T x^i)}^{\text{sigmoid}} - (1-y^i) \log \big(1 - \sigma(w^T x^i)\big) \Big) + C w_j^2$$

$$\xrightarrow[\text{گرادیان دسنت}]{\text{مشتق } w} \quad w_k = w_k - \alpha \Big(\sum_{i=1}^{n} \big(\sigma(w^T x^i) - y^i\big) x_j^i + 2 C w_j \underbrace{I(j=k)}_{\text{مشتر } j=k \text{ باشد ۱ وگرنه ۰}} \Big)$$

طبق لاس را مشاهده شده هنگامی که C خیلی بزرگ باشد ضریب کاهش w_j ($\alpha 2C$) زیاد شده و مدل w_j را به صفر نزدیک می‌کند.

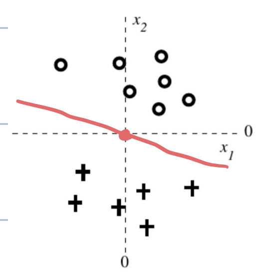
حالت ۱) $j = 0$:

در این مدل w_0 به صفر میل می‌کند. پس $w^T X \approx w_2 X_2 + w_1 X_1 + 0$

یعنی خط بین کلاس‌ها فیت شده در فضای X_1 و X_2 از مبدأ گذر می‌باشد.

حال چون خط با ارور ۰ مبدأ گذر داریم به همین فیت شده (چون psd است ارور مون به بهینه می‌رسه)

یعنی خطای ترین ۰

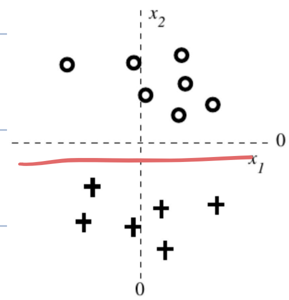


حالت دوم: $j=1$ ⇐ $w_1 \simeq 0$ ⇐ $w^T x \simeq w_2 x_2 + w_0$

به نوعی مستقل از ویژگی x_1 می‌شود.

در فضای x_1 و x_2 پس خطی عمود بر محور x_2 است. خط تمایز کننده برحسب x_2 است.

در مسئله فعلی با این ویژگی با تمایز ۱۰۰٪ داریم پس همچنان چون ارور صفر است خطاترین نمی‌کنیم



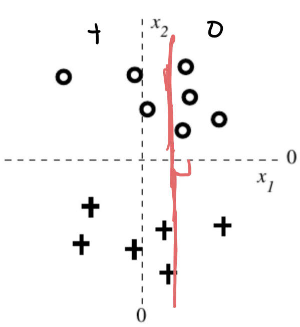
حالت سوم) $j=2$: $w_2 \simeq 0$ ⇐ $w^T x \simeq w_1 x_1 + w_0$

مشابه حالت دوم ولی این بار نسبت به متغیر x_2.

یعنی خطی عمود بر محور x_1 است و تمایز فقط برحسب x_1 است.

اما چنین خطی در فضا وجود ندارد که ارور ۰ به ما بدهد پس خطای تمرین در این مرحله بیشتر می‌شود.

چون فضا مقایسه فقط روی x_1 است انگار نقطه‌ها روی محور x_1 تصویر شوند برای همین خط عمود x_1 را گرفتم.



(خطی به صورت نقاط در شکل)

بهترین حالت را باید حالت‌های مختلف ببینیم. اگر خط قرمز بگیریم

$\frac{5+4}{5+4+3+1} = \frac{9}{13}$ اکیوریسی زیر را بیشترین داریم.

در کل خطا بیان > ۰ شد. افزایش یافت از دو تا قبلی. چون فقط x_1 تنها برای جداسازی اطلاعات کافی ندارد.

⑤

الف)

حالت ساده نشده :

$$P(y|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}\exp\{-\frac{1}{2\sigma^2}(y-\mu)^2\}$$

$$p(y;\eta)=b(y)\exp\{\eta^T T(y)-a(\eta)\}$$

$$\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\{-\frac{1}{2\sigma^2}(y^2-2y\mu+\mu^2)\}$$

$$\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2\sigma^2}y^2+\frac{y\mu}{\sigma^2}\}\exp\{\frac{-\mu^2}{2\sigma^2}-\ln(\sigma)\}$$

$$T(y)=\begin{pmatrix} y \\ y^2 \end{pmatrix} \qquad b(y)=\frac{1}{\sqrt{2\pi}}$$

$$\eta=\begin{pmatrix} \frac{\mu}{\sigma^2} \\ \frac{-1}{2\sigma^2} \end{pmatrix} \qquad a(\eta)=\frac{\mu^2}{2\sigma^2}+\ln(\sigma)=-\frac{\eta_1^2}{4\eta_2}-\frac{1}{2}\ln(-2\eta_2)$$

حالت ساده شده :

σ غیر متغیر ← فرقی نداره چند باشه 1 گرفتیمش

$$\Rightarrow \quad P(y;\mu)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}y^2)\exp(\mu y-\frac{1}{2}\mu^2)$$

$$\Rightarrow \quad \eta=\mu, \quad T(y)=y, \quad a(\eta)=\frac{\mu^2}{2}=\frac{\eta^2}{2} \qquad b(y)=\frac{1}{\sqrt{2\pi}}\exp(-\frac{y^2}{2})$$

ب)

ابتدا لم های را اثبات می‌کنیم.

* این ها را خواص اولیه می نامم.

قرار دهید $c(y)=e^{b(y)}$

اولی :

چون توزیع $\int P_\eta(y)\,dy = 1$ داریم

$$\xrightarrow{\nabla_\eta} \int \exp(\eta\cdot T(y) - a(\eta) + c(y))\big(T(y) - \nabla_\eta a(\eta)\big)\,dy = 0$$

$$\Rightarrow \int \exp(\eta\cdot T(y) - a(\eta) + c(y))\,T(y)\,dy = \int \exp(\eta\cdot T(y) - a(\eta) + c(y))\,\nabla_\eta a(\eta)\,dy$$

$$= \nabla_\eta a(\eta) \underbrace{\int \exp(\eta\cdot T(y) - a(\eta) + c(y))\,dy}_{1} = \nabla_\eta a(\eta)$$

$$\Rightarrow \nabla_\eta a(\eta) = \int \exp(\eta\cdot T(y) - a(\eta) + c(y))\,T(y)\,dy =$$

$$\int P_\eta(y)\,T(y)\,dy = E(T(y)) \qquad y \sim P_\eta(y)$$

اولی ادعا ثابت شد

دومی :

$$\frac{\partial a}{\partial \eta_i} = \int \exp(\eta\cdot T(y) - a(\eta) + c(y))\,T_i(y)\,dy$$

یک پرش صفحه بچینیم.

$$\frac{\partial^2 \alpha}{\partial \eta_i \partial \eta_j} = \int \exp(\eta \cdot T(y) - \alpha(\eta) + c(y)) \left(T_j(y) - \frac{\partial \alpha(\eta)}{\partial \eta_j}\right) T_i(y)\, dy$$

$$= \int \exp(\eta \cdot T(y) - \alpha(\eta) + c(y))\, T_i(y) T_j(y)\, dy - \frac{\partial \alpha(\eta)}{\partial \eta_j} \int \exp(\eta \cdot T(y) - \alpha(\eta))\, T_i(y)\, dy$$

$$= E(T_i(y) T_j(y)) - E(T_i(y)) E(T_j(y)) = \operatorname{cov}(T_i(y), T_j(y))$$

$$\Rightarrow D_\eta^2(\alpha) = \operatorname{Cov}(T(y))$$

اینجا چند تا لم اثبات می‌کنم برا خودم

لم ۱) درحالت کلی $\frac{\partial A}{\partial \eta^T} = E(T(X))$

$$\frac{\partial A}{\partial \eta^T} = \frac{\int T(u)\, e^{\eta^T T(u)}\, h(x)\, \nu(dx)}{\int T(x)\, e^{\eta^T T(x) - A(\eta)}\, h(x)\, \nu(dx)} = E(T(X))$$

لم ۲) $\frac{\partial A}{\partial \eta \partial \eta^T} = \operatorname{Var}(T(X))$

$$\frac{\partial^2 A}{\partial \eta \partial \eta^T} = \int T(x) \left(T(x) - \frac{\partial}{\partial \eta^T} A(\eta)\right)^T e^{\eta^T T(x) - A(\eta)}\, h(x)\, \nu(dx) =$$

$$\int T(x) \left(T(x) - E(T(x))\right)^T e^{\eta^T T(x) - A(\eta)}\, h(x)\, \nu(dx) = E(T(x) T(x)^T) -$$

$$E T(x)\, E T(x)^T = \operatorname{Var} T(X)$$

داریم در حالت گاوسیان که (۳م

$$\frac{\partial a(\eta)}{\partial \eta_1} = \frac{-2\eta_1}{4\eta_2} = \frac{-\eta_1}{2\eta_2} = \frac{-\frac{\mu}{\sigma^2}}{-\frac{1}{\sigma^2}} = \mu$$

(4م

$$\frac{\partial^2 a(\eta)}{\partial \eta_1^2} = \frac{\partial}{\partial \eta_1}\left(\frac{-\eta_1}{2\eta_2}\right) = -\frac{1}{2\eta_2} = -\frac{1}{2\times\left(-\frac{1}{2\sigma^2}\right)} = \sigma^2$$

حل ساده شده صورت آموزش تخته است. لم‌ها رو اول اون در تا اثبات کردم.

$$l(\theta) = -\sum_{i=1}^{m} \log b(y_i) + x_i^T \theta T(y_i) - a(\theta^T x_i)$$

$$\frac{\partial l}{\partial \theta} = -\sum_{i=1}^{m} T(y_i) x_i - \frac{\partial a}{\partial \theta}(\theta^T x_i) x_i$$

قبلاً ثابت کردم دو لم ماکسیمم مراحلش رو.

$$\frac{\partial l}{\partial \theta} = -\sum_{i=1}^{m} \left(T(y_i) - \frac{\partial a}{\partial \theta}(\eta)\right) x_i$$

$$\frac{\partial^2 l}{\partial \theta_j \partial \theta_k} = -\frac{\partial}{\partial \theta_j} \sum_{i=1}^{m} \left(T(y^i) - \frac{\partial a}{\partial \theta_k}(\eta)\right) x_k^i =$$

$$-\sum_{i=1}^{m}\left(-\frac{\partial a}{\partial \theta_j \partial \theta_k}(\eta) \frac{\partial \eta}{\partial \theta_j} x_k^i + 0\right) = \sum_{i=1}^{m} \frac{\partial a}{\partial \theta_j \partial \theta_k}(\eta) x_j^i x_k^i = H_{\text{هسین}}$$

$$\frac{\partial a}{\partial \theta_j} = \mu$$ پس $\nabla a = E(y)$ طبق لمی که ثابت کردم   طبق تخته ارائه شده:

$$\Rightarrow \frac{\partial a}{\partial \theta_j \partial \theta_k} = \frac{\partial \mu}{\partial \theta_k} \Rightarrow H = X^T W X \qquad W = \begin{bmatrix} \frac{\partial \mu}{\partial \theta_1} & \cdots & \vdots \\ \vdots & \ddots & \\ \vdots & \cdots & \end{bmatrix}$$

cov

حالت غیرساده شده

یک سری خواص که از قبل داریم :

① $\nabla a(\eta) = E(T(y))$

② $\nabla\nabla^T a(\eta) = Cov(T(y))$

حالت کامل کلی :

پایین‌تر مسئله را که حالت خاص اینه هم ثابت کردم

$$l = \sum_{i=1}^{m}\left[\log b(y^i) + \eta^T T(y^i) - a(\eta)\right]$$

$\eta = \theta^T x$

$T = (T_1(y), T_2(y), \dots)$ *

$$\nabla_{\theta_j} l = \sum_{i=1}^{m} x^i \left(T_j(y^i) - E(T_j(y^i)|\eta)\right)$$

از ①

$$\nabla_{\theta_j}\nabla_{\theta_k}^T l = \sum_{i=1}^{m} x^i x^{Ti} \left(-Cov[T_{j,k}(y^i)|\eta] + \delta_{jk}\eta_j''\left(T_k(y^i) - E[T_k(y^i)|\eta]\right)\right)$$

از ②

ترها معمولا حذف می‌شه

* حالت کمی ساده‌تر: در حالت $T(y)=y$ داریم که ماتریس :

$f(E(y)) = \eta$

هر خانواده متفاوته

$$\nabla_\theta l = X^T\left(y - f^{-1}(X\theta)\right) \rightsquigarrow \nabla_\theta\nabla_\theta^T l = -X^T W X$$

ماتریس واریانس‌هاست

$W_{ii} = var[y^i|\eta]$

مسئله :

$$l = -\sum_{i=1}^{m} \log b(y^i) + \eta^T T(y^i) - a(\eta)$$

$$\nabla_{\theta_j} l = -\sum_{i=1}^{m} x^{i^T} T_j(y^i) - \frac{\partial}{\partial \theta_j} a(\eta)$$

$\theta : r \times 2$

$X : r \times n$

$$\eta^T T(y^i) = x^{i^T} \theta \, T(y^i) = (X_1^1, \ldots, X_r^1) \begin{pmatrix} \theta_1^1 & \theta_1^2 \\ \vdots & \vdots \\ \theta_r^1 & \theta_r^2 \end{pmatrix} \begin{pmatrix} T_1 \\ T_2 \end{pmatrix}$$

$$= \left(X_1^1 \theta_1^1 + \cdots + X_r^1 \theta_r^1 ,\ X_1^1 \theta_1^2 + \cdots + X_r^1 \theta_r^2 \right) \begin{pmatrix} T_1 \\ T_2 \end{pmatrix}$$

$$= X_1^1 \theta_1^1 T_1 + \cdots + X_r^1 \theta_r^1 T_1 + X_1^1 \theta_1^2 T_2 + \cdots + X_r^1 \theta_r^2 T_2$$

$$\Rightarrow \frac{\partial \eta^T T(y^i)}{\partial \theta_j} = x^i T_j$$

$$\Rightarrow \nabla_{\theta_j} l = -\sum_{i=1}^{m} \left(x^i T_j(y^i) - \frac{\partial a(\eta)}{\partial \eta} \cdot \frac{\partial \eta}{\partial \theta_i} \right)$$

$$= -\sum_{i=1}^{m} x^i \left(T_j(y^i) - E(T_j(y^i)) \right)$$

در اینجا:

$$\frac{\partial l}{\partial \theta_1} = -\sum_{i=1}^{m} x^i (y^i - \mu)$$

$$\hookrightarrow E(T_1(y^i)) = E(y^i)$$

مشابه:

$$\frac{\partial l}{\partial \theta_2} = -\sum_{i=1}^{m} x^i \left(y^{i^2} - E(y^{i^2}) \right)$$

حال درباره مشتق ۲ گیریم :

$$\frac{\partial l}{\partial \theta_j \partial \theta_k} = +\sum_{i=1}^{m} x^i x^{iT} \operatorname{cov}\left(T_j(y^i), T_k(y^i)\right) \rightarrow H$$

$$\uparrow$$

$$\frac{\partial a}{\partial \theta_j \partial \theta_k} = \frac{\partial}{\partial \theta_j}\left(\frac{\partial a(\eta)}{\partial \eta_k} \cdot \frac{\partial \eta_k}{\partial \theta_k}\right) = \frac{\partial}{\partial \theta_j}\left(\frac{\partial a(\eta)}{\partial \eta_k}\right) \frac{\partial \eta_k}{\partial \theta_k} + 0$$

$$= \underbrace{\frac{\partial a(\eta)}{\partial \eta_j \partial \eta_k}}_{\operatorname{cov}} \frac{\partial \eta_j}{\partial \theta_j} \frac{\partial \eta_k}{\partial \theta_k} \rightarrow \operatorname{cov}\left(T_j(y^i), T_k(y^i)\right) x^i \cdot x^{iT}$$

منفی‌ها رو در هم زدیم $\rightarrow (-1) \times (-1) = 1$

$$\Rightarrow H = X^T S X$$

$$S = \operatorname{Cov}[T(y)]$$

همه نسبت به قبلاً بر حسب کرد.

حال کوواریانس بر حسب η_i هاهم نوشت.

$$E\left[\left(T_1(y) - E(T_1(y))\right)\left(T_2(y) - E(T_2(y))\right)\right]$$

$$= E\left[(y-\mu)\left(y^2 - E(y^2)\right)\right] = E\left((y-\mu)(y^2 - \mu^2 - \sigma^2)\right)$$

$$E(y^2) = E(y)^2 + E((y-E(y))^2) \Rightarrow$$

$$\Rightarrow E(y^3 - y\mu^2 - y\sigma^2 - \mu y^2 + \mu^3 + \mu\sigma^2)$$

$$= E(y^3) - \cancel{\mu^3} - \cancel{\mu\sigma^2} - \mu(\sigma^2+\mu^2) + \cancel{\mu^3} + \cancel{\mu\sigma^2} = E(y^3) - \mu(\sigma^2+\mu^2)$$

$$Var(T_1(y)) = E((y-E(y))^2) = \sigma^2$$

$$Var(T_2(y)) = E((y^2 - E(y^2))^2) = E((y^2-\mu^2-\sigma^2)^2) =$$

$$E(y^4) + \cancel{\mu^4} + \cancel{\sigma^4} - 2\mu^2(\sigma^2+\mu^2) - 2\sigma^2(\mu^2+\sigma^2) + 2\mu^2\sigma^2 =$$

$$E(y^4) - \mu^4 - \sigma^2 - 2\mu^2\sigma^2 = E(y^4) - (\mu^2+\sigma^2)^2$$

حالا که μ و σ اینا برحسب η_1 و η_2 درمیاد. X و X^T هم که از قبل ها داریم.

$$\eta = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ \frac{-1}{2\sigma^2} \end{pmatrix} \Rightarrow \frac{\eta_1}{-2\eta_2} = \mu$$

$$-\frac{1}{2\eta_2} = \sigma^2 \longrightarrow$$ برحسب η_1 و η_2 هست را داریم

$\longleftarrow$ برحسب η_1 ، η_2 ، X ها $X^T cov(T) X$ درومد :)