CS 229T

Ann He

July 2021

Linear algebra

Dual norm of L_1 norm

L_1 norm is

$$\|v\|_1 = \sum_{i=1}^{n} |v_i| \tag{1}$$

And by definition the dual is

$$\|v\|_* = \sup_{\|w\|\leq 1} v \cdot w \tag{2}$$

The dual norm tells us to "iterate" over all bounded norm vectors w and find the one with the largest dot product with the argument v. Since the dot product is linear, this is equivalent to placing all of the weight from our norm 1 vector on the largest magnitude element of v. So the dual norm returns the absolute value of the largest magnitude element of v. I think this is the 0-norm?

Trace is sum of singular values

The nuclear norm of a matrix A is

$$^{n}_{i=1}|\sigma_i(A)| \tag{3}$$

Where $\sigma_1(A), ..., \sigma_n(A)$ are the singular values of A

Show that the nuclear norm of a symmetric positive semidefinite matrix A is equal to its trace

$$\begin{aligned} tr(A) &= tr(PDP^{-1}) \\ &= tr(P^{-1}PD) \\ &= tr(D) \\ &= \sum_{i=1}^{n} \sigma_i(A) \\ &= \sum_{i=1}^{n} |\sigma_i(A)| \end{aligned}$$

When A is symmetric positive semidefinite matrix, the SVD puts its singular values on the diagonal.

Trace is bounded by nuclear norm

Subgradients of loss functions

Squared loss:

$$l(w; x, y) = \frac{1}{2}(y - w \cdot x)^2 \tag{4}$$

Hinge loss:

$$l(w; x, y) = max\{1 - yw \cdot x, 0\} \tag{5}$$

Convexity of loss functions

Squared

Let $f(w, x, y) = y - w \cdot x$, and
Let $g(z) = \frac{1}{2}z^2$

Then f is affine and therefore convex and g is convex because the square function is. By compositioinality of convexity, $g(f)$ is convex.

Hinge

Subgradients of loss functions

For the squared loss, we have $\partial f(w) = \{\nabla f(w)\}$. And

$$\begin{aligned}\nabla f(w) &= \nabla_w \frac{1}{2}(y - w \cdot x)^2 \\ &= (y - w \cdot x) \cdot x\end{aligned}$$

Subgradient of Hinge loss

$$\frac{\partial}{\partial w} 1 - yw \cdot x = \begin{cases} yx, & \text{if } 1 - yw \cdot x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Bound subgradients

Ok, let's try and bound the squared loss first...

$$\|(y - w \cdot x) \cdot x\| = |y - w \cdot x| \, \|x\|$$

Let's analyze $|y - w \cdot x|$

$$|y - w \cdot x| = max \begin{cases} |y| - |w \cdot x| \\ |w \cdot x| - |y| \end{cases}$$

By Cauchy-Schwarz, we have

$$|w \cdot x| \leq \|w\|_2^2 \, \|x\|_2^2 \tag{6}$$

And by our assumptions we have

$$|y| \leq 1 \tag{7}$$

So we can bound the squared loss subgradient as

$$\|w\|_2^2 \leq |1 - B^2C^2| \cdot C \tag{8}$$

For the hinge loss subgradient, we only have to consider the case where $1 - yw \cdot x \geq 0$. Then

$$\begin{aligned} \|g\|_2 &= \|yx\|_2 \\ &= |y| \, \|x\|_2 \\ &= C \end{aligned}$$

Probability bounds

Independent tail bound

We can't use the naive Union Bound on this because that would give us an upper bound and we are seeking a lower bound. However, the internet https://www.probabilitycourse.com/chapter6/6_2_1_union_bound_and_exten.php tells us that there are generalizations of the Union Bound based on the Inclusion-Exclusion principle.

Generalization of the Union Bound: Bonferroni Inequalities

For any events $A_1, A_2, ..., A_n$, we have:

$$\begin{aligned} &P\big(\cup_{i=1}^n A_i\big) \leq \sum_{i=1}^n P(A_i) \\ &P\big(\cup_{i=1}^n A_i\big) \geq \sum_{i=1}^n P(A_i) - \sum_{i<j} P(A_i \cap A_j) \\ &P\big(\cup_{i=1}^n A_i\big) \leq \sum_{i=1}^n P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) \end{aligned}$$

To begin, let A_i be the event that $f(x_i)$ is an error (i.e. example i is classified incorrectly by

f. We want: $P(\cup_{i=1}^n A_i) \geq 1-\delta$. Let's try to use a Bonferroni Inequality...

$$\begin{aligned}
P(\cup_{i=1}^n A_i) &\geq \sum_{i=1}^n P(A_i) - \sum_{i<j} P(A_i \cap A_j) \\
&= n\alpha - \sum_{i<j} P(A_i \cap A_j) \\
&= n\alpha - \sum_{i<j} P(A_i)P(A_j) \\
&= n\alpha - \frac{n(n-1)}{2}\alpha^2 \\
&= 1-\delta
\end{aligned}$$

Solving for n we get

$$n = \frac{(\alpha + \frac{1}{2}\alpha^2) + -(-\alpha - \frac{1}{2}\alpha^2)^2 - 4(1-\delta)(\frac{1}{2}\alpha^2)}{\alpha^2} \tag{9}$$

To-do: Check conditions for positive n

Asymptotics

Gaussian tail bound

Using a simple application of Markov's inequality...

$$\begin{aligned}
P[Z - \mathbb{E}Z \geq t] &= P[\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i \geq t] \\
&\leq \frac{\mathbb{E}[exp(\lambda(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i))]}{e^{\lambda t}} \\
\log P[Z - \mathbb{E}Z \geq t] &\leq \log \mathbb{E}[exp(\lambda(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i))] - \log e^{\lambda t}
\end{aligned}$$

Expanding...

$$\begin{aligned}
\log \mathbb{E}[exp(\lambda(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i))] &= \log \mathbb{E}[e^{\lambda(X_1-\mu_1)}]\mathbb{E}[e^{\lambda(X_2-\mu_2)}]...\mathbb{E}[e^{\lambda(X_n-\mu_n)}] \\
&= \frac{\lambda^2 \sum_{i=1}^n \sigma_i^2}{2}
\end{aligned}$$

So we have

$$\log P[Z - \mathbb{E}Z \geq t] \leq -sup_\lambda\{\lambda t - \frac{\lambda^2 \sum_{i=1}^n \sigma_i^2}{2}\} \tag{10}$$

Solving the derivative, we get

$$P[Z - \mathbb{E}Z \geq t] \leq e^{-t^2/(2\sum_{i=1}^n \sigma_i^2)} \tag{11}$$

Moments

Variance algebra

$$
\begin{aligned}
Var(aX_1 + bX_2) &= a^2Var(X_1) + b^2Var(X_2) + 2abCov(X_1, X_2) \\
&= a^2\sigma_1^2 + b^2\sigma_2^2
\end{aligned}
$$

Decomposition lemma

Lemma 0.1. *Decomposition lemma. Let Z be a real-valued random variable and a be a real constant. Then,* $\mathbb{E}[(Z-a)^2] = (\mathbb{E}[Z]-a)^2 + Var(Z)$

Proof.

$$
\begin{aligned}
\mathbb{E}[(Z-a)^2] &= \mathbb{E}[Z^2 - 2aZ + a^2] \\
&= \mathbb{E}[Z^2] - 2a\mathbb{E}[Z] + a^2
\end{aligned}
$$

Recall..

$$Var(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 \tag{12}$$

and

$$\mathbb{E}[Z^2] = Var(Z) + \mathbb{E}[Z]^2 \tag{13}$$

So

$$
\begin{aligned}
\mathbb{E}[(Z-a)^2] &= \mathbb{E}[Z^2 - 2aZ + a^2] \\
&= \mathbb{E}[Z^2] - 2a\mathbb{E}[Z] + a^2 \\
&= Var(Z) + \mathbb{E}[Z]^2 - 2a\mathbb{E}[Z] + a^2 \\
&= Var(Z) + (\mathbb{E}[Z] - a)^2
\end{aligned}
$$

Moments of mixture models

$$
\begin{aligned}
\mathbb{E}[x_1x_2^T] &= \sum_{i=1}^{K} \mathbb{E}[x_1|h]\mathbb{E}[x_2|h]^T Pr(h=i) \\
&= \sum_{i=1}^{K} \pi_i\mu_i\mu_i^T \\
&= M\mu M^T
\end{aligned}
$$

Exponential families

Moment generating properties of exponential families

Derivatives of log partition function

$$
\begin{aligned}
\nabla_\theta A(\theta) &= \nabla_\theta \log \sum_{x\in\mathcal{X}} exp\{\theta\cdot\phi(x)\} \\
&= \frac{1}{\sum_{x\in\mathcal{X}} exp\{\theta\cdot\phi(x)\}} \nabla_\theta \sum_{x\in\mathcal{X}} exp\{\theta\cdot\phi(x)\} \\
&= \frac{1}{\sum_{x\in\mathcal{X}} exp\{\theta\cdot\phi(x)\}} \sum_{x\in\mathcal{X}} \nabla_\theta exp\{\theta\cdot\phi(x)\} \\
&= \frac{1}{\sum_{x\in\mathcal{X}} exp\{\theta\cdot\phi(x)\}} \sum_{x\in\mathcal{X}} exp\{\theta\cdot\phi(x)\}\cdot\phi(x)
\end{aligned}
$$

Now we argue that $\nabla A(\theta)$ is convex, as a function of θ.

First note that $\frac{1}{\sum_{x\in\mathcal{X}} exp\{\theta\cdot\phi(x)\}}$ is convex. It is the composition of $f(z) = z^{-1}$, which is convex because power functions x^α with $\alpha \geq 1$ and $\alpha \leq 0$ are convex, and the sum (convex) of exponential (convex) of dot product (affine and therefore convex). (see https://web.stanford.edu/class/ee364a/lectures/functions.pdf). And $\sum_{x\in\mathcal{X}} exp\{\theta\cdot\phi(x)\}\cdot\phi(x)$ can be argued similarly. Finally, the product of convex functions is also convex.

Now we take the second derivative.

$$
\nabla^2_\theta A(\theta) = \nabla_\theta \sum_{x\in\mathcal{X}} \frac{exp\{\theta\cdot\phi(x)\}}{\sum_{x\in\mathcal{X}} exp\{\theta\cdot\phi(x)\}}\cdot\phi(x)
$$

To make this more explicit let n be the cardinality of $\mathcal{X}$. Then we have

$$
\nabla^2_\theta A(\theta) = \nabla_\theta \sum_{i=1}^{n} \frac{exp\{\theta\cdot\phi(x_i)\}}{\sum_{j=1}^{n} exp\{\theta\cdot\phi(x_j)\}}\cdot\phi(x_i)
$$

Let $s_i = \frac{exp\{\theta\cdot\phi(x_i)\}}{\sum_{j=1}^{n} exp\{\theta\cdot\phi(x_j)\}}\cdot\phi(x_i)$. Now, rewrite the expression as

$$
\begin{aligned}
\nabla^2_\theta A(\theta) &= \nabla_\theta \sum_{i=1}^{n} s_i(\theta)\cdot\phi(x_i) \\
&= \sum_{i=1}^{n} \nabla_\theta s_i(\theta)\cdot\phi(x_i)
\end{aligned}
$$

Ok, now let's try a trick...

$$\nabla_\theta \log s_i(\theta) = \frac{1}{s_i(\theta)} \nabla_\theta s_i(\theta)$$

Rearranging, we get

$$\nabla_\theta s_i(\theta) = s_i \cdot \nabla_\theta \log s_i(\theta)$$

Wait.... Let's try to log the whole expression so we can get rid of the product as well as the quotient rule...

$$\nabla_\theta \big(\log s_i(\theta) \cdot \phi(x_i) \big) = \frac{1}{s_i(\theta) \cdot \phi(x_i)} \nabla_\theta \big(s_i(\theta) \cdot \phi(x_i) \big)$$

So...

$$\nabla_\theta \big(s_i(\theta) \cdot \phi(x_i) \big) = s_i(\theta) \cdot \phi(x_i) \nabla_\theta \big(\log s_i(\theta) \cdot \phi(x_i) \big)$$

Yeah, I think that's better... So,

$$\begin{aligned}
\nabla_\theta^2 A(\theta) &= \sum_{i=1}^{n} s_i(\theta) \cdot \phi(x_i) \nabla_\theta \big(\log \big[s_i(\theta) \cdot \phi(x_i) \big] \big) \\
&= \sum_{i=1}^{n} s_i(\theta) \cdot \phi(x_i) \nabla_\theta \big(\log s_i(\theta) + \log \phi(x_i) \big) \\
&= \sum_{i=1}^{n} s_i(\theta) \cdot \phi(x_i) \nabla_\theta \big(\log exp\{\theta \cdot \phi(x_i)\} - \log \sum_{j=1}^{n} exp\{\theta \cdot \phi(x_j)\} + \log \phi(x_i) \big)
\end{aligned}$$

I think the rest of these calculations should be pretty straightforward... To-do: Later

Optimization with the entropy

$$\begin{aligned}
\mathcal{L}(p, \lambda) &= f(p) - \lambda(g(p) - 1) \\
&= H(p) - \lambda(\sum_{x \in \mathcal{X}} p(s) - 1)
\end{aligned}$$

Taking derivatives with respect to p and λ...

$$\begin{aligned}
\nabla_{pk} \mathcal{L}(p, \lambda) &= -\log p_k(x) - 1 - \lambda \\
&= 0
\end{aligned}$$

and

$$\begin{aligned}\nabla_\lambda \mathcal{L}(p, \lambda) &= -\sum_{x \in \mathcal{X}} p(s) + 1 \\ &= 0\end{aligned}$$

If p is d-dimensional, then there are $d+1$ variables and $d+1$ equations. Let's solve for the unknowns

$$\begin{aligned}\log p_k(x) &= -1 - \lambda \\ p_k(x) &= e^{-1-\lambda}\end{aligned}$$

Let's solve for λ

$$\begin{aligned}\sum_{x \in \mathcal{X}} p(x) &= 1 \\ \sum_{x \in \mathcal{X}} e^{-1-\lambda} &= 1 \\ de^{-1-\lambda} &= 1 \\ \lambda &= -1 - \log(1/d) \\ \lambda &= \log d - 1\end{aligned}$$

Side note: This is a generic technique called the principle of maximum entropy https://sgfin.github.io/2017/03/16/Deriving-probability-distributions-using-the-Principle-of-Maximum-Entropy/