CS 419M Introduction to Machine Learning **Spring 2021-22**

Lecture 10: Duality in Loss Functions

Lecturer: Abir De *Scribe: Group 1*

10.1 Bias and Regularization

Consider the original loss function,

$$F_D(w) = \sum_{i \in D} [l(w^T x_i, y_i) + \lambda ||w||^2] \tag{10.1}$$

Now suppose we add a bias here,

$$F_D(w, b) = \sum_{i \in D} [l(w^T x_i + b, y_i) + \lambda ||w||^2] \tag{10.2}$$

We notice in the above expression that b is not regularized. Usually we do not regularize the bias, because there is no need for it. The understanding behind this is, that if we don't provide any data, then we get w to be 0, and we can just keep the prediction to be b, whatever be the value of b. In other words, if we don't provide any data, there is no point of assigning weights to features, and the offset can be anything, it doesn't have any significant meaning.
Alternate explanation: Let us assume that b is a part of the regularization term. In order to minimize $F(w, b)$, the model will reduce the value of all parameters/weights including b. Therefore, after regularization, b can take either negative or positive values but it is assured that these values will be closer to 0. This will cause the line of best fit to move towards the origin and end up in a similar situation as fixing the bias to 0 thus leading to underfitting.

10.1.1 Implications of unregularized bias

Now as long as, b is not regularized, the loss function need not be stable. But we want $F_D(w, b)$ to be strictly convex w.r.t b as well.

Now, if we don't regularize F w.r.t b, then minimum eigenvalue of $\frac{\partial^2 F}{\partial b^2}$ is equal to 0, implying that $F(w, b)$ is not strictly convex. If you assume that the loss function is such that it is strictly convex without even regularizing, then it will be stable. Therefore the case where the b is not regularized, can be strictly convex, but need not always be strictly convex.

For example, consider the following SVM loss,

$$l(w^T x_i + b, y) = (1 - y(w^T x_i + b))_+$$

If we differentiating once, we get $\frac{\partial l}{\partial b} = 0$ or y.
Now we get $\frac{\partial^2 l}{\partial b^2} = 0$ (except at the junction)

Questions: If we include the bias in the w itself, then it would be stable, right?
Answer: Yes
Questions: Suppose we also want to regularize b, then how would this $F_D(w, b)$ look like?
Answer: Just add the term λb^2. Here the value of λ can either be same as that of w or different.

Let us consider another case. How a SVM looks like in an unconstrained space.

$$\min_{w,b} \lambda||w||^2 + \sum_{i \in D} l(w^T x_i + b, y_i)$$

For the time being, assume D is fixed, here we are not discussing stability. Here we have taken a single fixed λ. Also, in the support vector machine context, hinge loss is used. Now, by replacing λ with 1/c, we get the following,

$$\min_{w,b} ||w||^2 + c \sum_{i \in D} (1 - (w^T x_i + b) y_i)_+ \tag{10.3}$$

This is equivalent to,

$$\min_{w,b,\zeta \geq 0} ||w||^2 + \sum_{i \in D} c\zeta_i \quad ; \quad \text{Here, } y_i(w^T x_i + b) \geq 1 - \zeta_i \forall i \in D \tag{10.4}$$

Now this is a constrained problem, how do we solve this problem? To find the minimum in a constraint space, we go into the unconstrained space and convert this problem into the unconstrained dual. Now, what is the dual, it is same as the following,

$$\max_{\alpha \geq 0, \beta \geq 0} \min_{w,b,\zeta \geq 0} ||w||^2 + c \sum_{i \in D} \zeta_i + \alpha_i (1 - \zeta_i - y_i(w^T x_i + b)) - \beta_i \zeta_i \tag{10.5}$$

Why are we writing this? Suppose you are given an optimization problem like the following,

$$\min_x f(x) \quad ; \quad \text{such that } g(x) \leq 0$$

The above problem is same(at least in the convex space) as writing the following,

$$\max_{\alpha \geq 0} \min_x (f(x) + \alpha g(x)) \tag{10.6}$$

We can think that we are trying to penalize it with respect to g(x) in the above expression. There are rigorous proofs of the same, but the intuition is that we can think α as a regularizer, because there is a g(x) which we want to be less than zero so if it is positive we are trying to minimize it, so that is why you are taking like alpha greater than zero. Therefore, any constrained optimization problem where you want to minimize the function with respect to x such that g(x) is less than equal to 0 this is equivalent to the above expression

Q: Why are we maximizing it with respect to alpha?
Answer: The intuition foes like, we can see that when you minimize $(f(x) + \alpha g(x))$ w.r.t to x, then we get a function of alpha. And one can show that this function of alpha actually looks like a quadratic with negative leading coefficient(concave function) w.r.t α. And f(x) is like a quadratic

equation with positive leading coefficient(convex function) w.r.t x. So they meet at one point. Rigorous proof is beyond the scope of this class.

Q: If we are saying that the two curves intersect then the function needs to be strongly dual, right?
Answer: Yes, I have assumed that. For the strongly dual, it is the case. It typically looks like this if it is strongly dual.

Exercises: Now, is loss function [10.3] convex?
Answer: For positive c, the entire function is convex.

Now, suppose let us consider another loss. Currently it is hinge loss, so instead of hinge loss you consider something else. Consider the following expression,

$$\max_{\alpha\geq 0,\beta\geq 0}\min_{w,b,\zeta\geq 0} ||w||^2 + c\sum_{i\in D}\zeta_i + \alpha_i(l(w^T x_i, y_i) - \zeta_i) - \beta_i\zeta_i \tag{10.7}$$

Let $F(w,b) = ||w||^2 + c\sum_{i\in D}\zeta_i + \alpha_i(l(w^T x_i, y_i) - \zeta_i) - \beta_i\zeta_i$

$$\frac{\partial F}{\partial w} = 2w - \sum_{i\in D}\alpha_i x_i y_i = 0 \tag{10.8}$$

$$\implies w = (\sum_{i\in D}\alpha_i x_i y_i)/2 \tag{10.9}$$

$$\frac{\partial F}{\partial b} = 0 \tag{10.10}$$

$$\implies \sum_{i\in D}\alpha_i y_i = 0 \tag{10.11}$$

$$\frac{\partial F}{\partial \zeta} = c - \alpha_i - \beta_i = 0 \tag{10.12}$$

$$\implies c = \alpha_i + \beta_i \tag{10.13}$$

Q:If we are optimising w.r.t w, b, ζ, how come we can satisfy last constraint?
Answer: We are only considering the cases where the minima would exist.

Now, satisfying these constraint, the optimisation problem is

$$\max_{\alpha\geq 0,\beta\geq 0}\sum_{i\in D}\alpha_i - \sum_{i,j\in D}\alpha_i\alpha_j x_i^T x_j y_i y_j \tag{10.14}$$

as $\beta_i, \alpha_i > 0$ and $c = \alpha_i + \beta_i \implies c \geq \alpha \geq 0$ The problem reduces to

$$\max_{c\geq\alpha\geq 0}\sum_{i\in D}\alpha_i - \sum_{i,j\in D}\alpha_i\alpha_j x_i^T x_j y_i y_j \tag{10.15}$$

Also we see $\sum_{i \in D} \alpha_i - \sum_{i,j \in D} \alpha_i \alpha_j x_i^T x_j y_i y_j$ is concave.

Applying Slater's condition, we see

$$\alpha_i((w^T x_i + b)y_i - 1 + \zeta_i) = 0 \tag{10.16}$$

$$\beta_i \zeta_i = 0 \tag{10.17}$$

If we assume $(w^T x_i + b)y_i - 1 > 0 \implies \zeta_i = 0 \implies \alpha_i = 0$

Alternatively, if we assume $(w^T x_i + b)y_i - 1 < 0$ and we have $\beta_i \zeta_i = 0, \implies \zeta_i > 0 \implies \beta_i = 0 \implies \alpha_i = c$

On the edge case $(w^T x_i + b)y_i - 1 = 0 \implies \alpha_i \zeta_i = 0 \implies \zeta_i = 0$ But we can't decide on α_i and β_i. But they will be constrained as $\alpha_i + \beta_i = 0$

10.2 Group Details and Individual Contribution

- 20D070060-Prateek Garg
- 190010061-Shashank Singh
- 200040054-Gautam Asodiya
- 200100054-Desai Utsav Manojkumar
- 200100045-Bheemrao Badiger