AGML, Sommersemester 2020 Vorlesung 13 (5 Seiten)

Regularisierung

Thomas Kesselheim Letzte Aktualisierung: 12. Juni 2020

In der letzten Vorlesung haben wir das Phänomen des Overfitting kennengelernt. Zu Erinnerung: Wir nehmen an, dass ein Lernalgorithmus eine Trainingsmenge von m Datenpunkt-/Label-Paare aus $X \times Y$ erhält und mithilfe von diesem Sample eine Hypothese $h_S \colon X \to Y$ finden soll, die Labels für Datenpunkte vorhersagen sollen. Beim Overfitting tritt es auf, dass die Hypothese „zu gut“ auf den Trainingsdaten ist und sich daher zu schlecht verallgemeinert. Eine gute Faustregel ist, dass man „einfachere“ Hypothesen verwenden sollte, um Overfitting zu vermeiden. Hierzu werden wir heute ein formales Argument führen.

Wir haben bereits die Definition eines stabilen Lernalgorithmus eingeführt.

Definition 13.1. *Sei $\delta \colon \mathbb{N} \to \mathbb{R}$. Ein Lernalgorithmus ist* universell δ-austauschstabil*, wenn für alle $m \in \mathbb{N}$, alle Mengen S von m Datenpunkt-/Label-Paaren, alle $i \in \{1, \dots, m\}$ und alle weiteren Datenpunkt-/Label-Paare z' gilt*

$$\ell(h_{S^i}, z_i) - \ell(h_S, z_i) \leq \delta(m) \ .$$

Hierbei ist $\ell(h, z)$ der *Loss* von Hypothese h auf $z \in X \times Y$. Dieser drückt aus, „wie falsch“ die Hypothese h auf z ist. Unsere Erkenntnis hinsichtlich Overfitting lässt sich knapp zusammenfassen als:

> Ein universell δ-austauschstabiler Lernalgorithmus mit $\delta(m) \to 0$ für $m \to \infty$ vermeidet Overfitting.

Heute werden mit *Regularisierung* einen grundsätzlichen Ansatz kennenlernen, der zu Stabilität führt. Anstatt eine Hypothese h_S zu wählen, sodass $L_S(h_S)$ minimiert wird, sollten „extreme“ Hypothesen vermieden werden.

1 Annahmen

Wir betrachten heute keine beliebigen Hypothesenklassen mehr, sondern treffen ein paar Annahmen. Zunächst einmal nehmen wir an, dass die Hypothesen in unsere Klasse $\mathcal{H}$ durch Vektoren $\mathbf{w} \in \mathbb{R}^n$ parametrisiert sind. Das heißt,

$$\mathcal{H} = \{h_{\mathbf{w}} \colon X \to Y \mid \mathbf{w} \in M\} \ ,$$

wobei $M \subseteq \mathbb{R}^n$ eine konvexe Menge ist. Ein typisches Beispiel sind lineare Klassifikatoren (hier ist $Y = \{-1, +1\}$)

$$h_{\mathbf{w}}(\mathbf{x}) = \begin{cases} +1 & \text{falls } \langle \mathbf{w}, \mathbf{x} \rangle \geq 0 \\ -1 & \text{sonst} \end{cases} \ .$$

Wie wir gesehen haben, können mittels Einbettungen in einen Feature Space auch andere Hypothesen so dargestellt werden.

Analog kann man lineare Regression darstellen (nun ist $Y = \mathbb{R}$) über

$$h_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle \ .$$

Für unsere Ergebnisse wird vollkommen unerheblich sein, wie die Hypothese $h_{\mathbf{w}}$ genau definiert ist. Wir nehmen lediglich an, dass die Loss-Funktionen konvex sind. Das heißt, dass $\mathbf{w} \mapsto \ell(h_{\mathbf{w}}, z)$ konvex ist für alle z.

Darüber hinaus nehmen wir an, dass sie ρ-Lipschitz sind. Das heißt, dass für alle $\mathbf{w}, \mathbf{w}' \in M$ und alle z

$$\ell(h_{\mathbf{w}}, z) - \ell(h_{\mathbf{w}'}, z) \leq \rho \|\mathbf{w} - \mathbf{w}'\| \ .$$

Beispiel 13.2. *Der 0/1 Loss ist nicht konvex. Entsprechend sind unsere heutigen Ergebnisse nicht anwendbar.*

Der Hinge Loss auf $z = (x, y)$ ist definiert als

$$\ell^{\text{hinge}}(h_{\mathbf{w}}, \mathbf{z}) = \max\{0, 1 - y\langle \mathbf{w}, \mathbf{x}\rangle\} \ .$$

Er ist $\|\mathbf{x}\|$-Lipschitz.

Der quadratische Loss (für Regression) ergibt sich zu

$$\ell^{\text{squared}}(h_{\mathbf{w}}, z) = (\langle \mathbf{w}, \mathbf{x}\rangle - y)^2 \ .$$

Er ist ρ-Lipschitz für $\rho = 2\|\mathbf{x}\|^2 \max_{\mathbf{w}\in M} \|\mathbf{w}\|$.

2 Starke Konvexität

Wir werden nun eine genauere Definition von Konvexität einführen, die zum Ausdruck bringt, wieviel deutlicher eine Funktion wächst als eine lineare Funktion. Dafür vergleichen wir sie mit einer quadratischen Funktion.

Definition 13.3. *Sei $\sigma \geq 0$. Eine Funktion $f\colon M \to \mathbb{R}$ heißt σ-stark konvex, wenn für alle $\mathbf{u}, \mathbf{v} \in M$ und alle $\lambda \in [0, 1]$ gilt*[1]

$$f(\lambda \mathbf{u} + (1-\lambda)\mathbf{v}) \leq \lambda f(\mathbf{u}) + (1-\lambda) f(\mathbf{v}) - \frac{\sigma}{2}\lambda(1-\lambda)\|\mathbf{u} - \mathbf{v}\|^2 \ .$$

Eine Funktion ist konvex genau dann, wenn sie 0-stark konvex ist.

Konvexität erfordert, dass die Funktion f jeweils unterhalb der Verbindungslinien auf dem Funktionsgraphen bleibt. Starke Konvexität mit $\sigma > 0$ fordert zusätzlich, dass sie unterhalb einer verbindenden Parabel bleibt. Das heißt, die Funktion muss „durchhängen“ (siehe Abbildung 1).

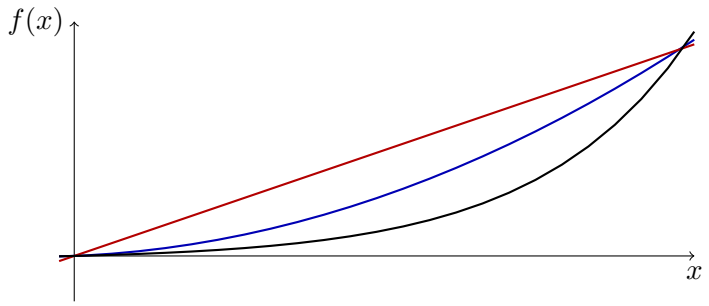

Abbildung 1: Eine stark konvexe Funktion in schwarz mit einer direkten Verbindunglinie zweier Punkt in rot und einer dazwischen liegenden Parabel in blau.

[1] Es mag etwas verwundern, dass der Faktor $\frac{\sigma}{2}$ ist und nicht σ. Auf diese Weise bleibt die Definition äquivalent mit anderen in der Literatur üblichen Formulierungen.

Beispiel 13.4. *Für jedes $\alpha \geq 0$, ist Funktion $f\colon \mathbb{R}^n \to \mathbb{R}$, $f(\mathbf{x}) = \alpha\|\mathbf{x}\|^2$ jeweils 2α-stark konvex.*

Für alle $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ und alle $\lambda \in [0,1]$ gilt

$$\begin{aligned}\|\lambda\mathbf{u} + (1-\lambda)\mathbf{v}\|^2 &= \sum_{i=1}^{n}(\lambda u_i + (1-\lambda)v_i)^2 = \sum_{i=1}^{n}(\lambda u_i)^2 + ((1-\lambda)v_i)^2 + 2\lambda u_i(1-\lambda)v_i \\ &= \lambda^2\|\mathbf{u}\|^2 + (1-\lambda)^2\|\mathbf{v}\|^2 + 2\lambda(1-\lambda)\langle\mathbf{u},\mathbf{v}\rangle \\ &= \lambda\|\mathbf{u}\|^2 - \lambda(1-\lambda)\|\mathbf{u}\|^2 + (1-\lambda)\|\mathbf{v}\|^2 - \lambda(1-\lambda)\|\mathbf{v}\|^2 + 2\lambda(1-\lambda)\langle\mathbf{u},\mathbf{v}\rangle \\ &= \lambda\|\mathbf{u}\|^2 + (1-\lambda)\|\mathbf{v}\|^2 - \lambda(1-\lambda)\|\mathbf{u}-\mathbf{v}\|^2 \ .\end{aligned}$$

Indem wir beide Seiten dieser Gleichung mit α multiplizieren, erhalten wir

$$f(\lambda\mathbf{u} + (1-\lambda)\mathbf{v}) = \lambda f(\mathbf{u}) + (1-\lambda)f(\mathbf{v}) - \frac{2\alpha}{2}\lambda(1-\lambda)\|\mathbf{u}-\mathbf{v}\|^2 \ .$$

Das heißt, die geforderte Ungleichung ist für $\sigma = 2\alpha$ sogar mit Gleichheit erfüllt.

Die Bedeutung von stark konvexen Funktionen zeigt sich im folgenden Lemma. Es sagt aus, dass wir in deutlicher Entfernung vom Minimum auch deutlich größere Funktionswerte beobachten.

Lemma 13.5. *Sei $f\colon M \to \mathbb{R}$ eine σ-stark konvexe Funktion. Sei $\mathbf{w} \in \arg\min_{\mathbf{v}\in M} f(\mathbf{v})$ ein Punkt, der f minimiert. Dann gilt für alle $\mathbf{u} \in M$*

$$f(\mathbf{u}) - f(\mathbf{w}) \geq \frac{\sigma}{2}\|\mathbf{u}-\mathbf{w}\|^2 \ .$$

Beweis. Wir betrachten die Verbindungslinie zwischen $\mathbf{u}$ und $\mathbf{w}$. Für alle $\lambda \in [0,1]$ haben wir gemäß starker Konvexität

$$f(\lambda\mathbf{u} + (1-\lambda)\mathbf{w}) \leq \lambda f(\mathbf{u}) + (1-\lambda)f(\mathbf{w}) - \frac{\sigma}{2}\lambda(1-\lambda)\|\mathbf{u}-\mathbf{w}\|^2 \ .$$

Gleichzeitig wird f durch $\mathbf{w}$ minimiert. Also

$$f(\lambda\mathbf{u} + (1-\lambda)\mathbf{w}) \geq f(\mathbf{w}) \ .$$

Somit gilt für alle $\lambda \in [0,1]$

$$\lambda f(\mathbf{u}) + (1-\lambda)f(\mathbf{w}) - \frac{\sigma}{2}\lambda(1-\lambda)\|\mathbf{u}-\mathbf{w}\|^2 \geq f(\mathbf{w}) \ .$$

Falls $\lambda > 0$ ist, ist dies äquivalent zu

$$f(\mathbf{u}) - f(\mathbf{w}) \geq \frac{\sigma}{2}(1-\lambda)\|\mathbf{u}-\mathbf{w}\|^2 \ .$$

Angenommen, es gilt nun

$$f(\mathbf{u}) - f(\mathbf{w}) < \frac{\sigma}{2}\|\mathbf{u}-\mathbf{w}\|^2 \ ,$$

dann müsste auch

$$f(\mathbf{u}) - f(\mathbf{w}) < c\frac{\sigma}{2}\|\mathbf{u}-\mathbf{w}\|^2$$

für irgendein $c < 1$ gelten. Dann könnten wir $\lambda = 1 - c$ wählen und würden einen Widerspruch erhalten. Also gilt das Lemma. □

Wir halten noch eine einfache Beobachtung fest, die sich durch Nachrechnen zeigen lässt.

Beobachtung 13.6. *Ist $f_1\colon M \to \mathbb{R}$ eine σ-stark konvexe Funktion, $f_2\colon M \to \mathbb{R}$ eine konvexe Funktion, dann ist $f_1 + f_2$ eine σ-stark konvexe Funktion.*

3 Stark konvexe Regularisierung führt zu Stabilität

Wir betrachten nun den Lernalgorithmus, der anstatt $\mathbf{w}$ zu finden, sodass $L_S(h_\mathbf{w})$ minimiert wird, eine *regularisierte* Zielfunktion $f(\mathbf{w}) = R(\mathbf{w}) + L_S(h_\mathbf{w})$ minimiert. Konkret ist in unserem Fall $R(\mathbf{w}) = \alpha\|\mathbf{w}\|^2$. Wie wir oben gesehen haben, ist R nun 2α-stark konvex und somit auch f.

Beispiel 13.7. *Für lineare Klassifikation mittels Hinge Loss ergibt sich genau das Soft-SVM-Problem*[2]*.*

Für Regression nennt sich die Vorgehensweise $\alpha\|\mathbf{w}\|^2 + L_S^{\text{squared}}(h_\mathbf{w})$ *zu minimieren* Ridge Regression.

Wir können nun zeigen, dass jeder Lernalgorithmus, der eine stark-konvexe Regularisierungsfunktion verwendet, stabil ist.

Satz 13.8. *Sind die Loss-Funktionen konvex und ρ-Lipschitz und ist die Regularisierungsfunktion σ-stark konvex, dann ist der Lernalgorithmus universell $\frac{2\rho^2}{m\sigma}$-austauschstabil.*

Es ist wichtig, dass $\delta(m) = \frac{2\rho^2}{m\sigma}$ gegen 0 konvergiert. Gemäß der Ergebnisse aus der letzten Vorlesung heißt das, dass der erwartete Verallgemeinerungsfehler verschwindet, wenn wir genügend Samples verwenden.

Beweis von Satz 13.8. Sei $\mathbf{w}^*$ der Vektor, der die Hypothese beschreibt, die der Lernalgorithmus auf S berechnet. Das heißt, $h_S = h_{\mathbf{w}^*}$. Analog sei $\mathbf{w}^i$ der entsprechende Vektor für die Lösung auf S^i.

Laut Definition minimiert $\mathbf{w}^*$ die Funktion $f(\mathbf{w}) := R(\mathbf{w}) + \frac{1}{m}\sum_{j=1}^m \ell(h_\mathbf{w}, z_j)$. Andererseits minimiert $\mathbf{w}^i$ die Funktion $f^i(\mathbf{w}) := R(\mathbf{w}) + \frac{1}{m}\sum_{j=1, j\neq i}^m \ell(h_\mathbf{w}, z_j) + \ell(h_\mathbf{w}, z')$.

Deshalb erhalten wir jeweils durch Anwendung von Lemma 13.5

$$f(\mathbf{w}^i) - f(\mathbf{w}^*) \geq \frac{\sigma}{2}\|\mathbf{w}^i - \mathbf{w}^*\|^2$$

und

$$f^i(\mathbf{w}^*) - f^i(\mathbf{w}^i) \geq \frac{\sigma}{2}\|\mathbf{w}^* - \mathbf{w}^i\|^2 \ .$$

In Kombination also

$$f(\mathbf{w}^i) - f(\mathbf{w}^*) + f^i(\mathbf{w}^*) - f^i(\mathbf{w}^i) \geq \sigma\|\mathbf{w}^i - \mathbf{w}^*\|^2$$

Wenn wir die Definitionen von f und f^i einsetzen, erhalten wir die äquivalente Ungleichung

$$\frac{1}{m}\ell(h_{\mathbf{w}^i}, z_i) - \frac{1}{m}\ell(h_{\mathbf{w}^i}, z') - \frac{1}{m}\ell(h_{\mathbf{w}^*}, z_i) + \frac{1}{m}\ell(h_{\mathbf{w}^*}, z') \geq \sigma\|\mathbf{w}^i - \mathbf{w}^*\|^2 \ .$$

Durch die Lipschitz-Bedingungen können wir abschätzen

$$\ell(h_{\mathbf{w}^i}, z_i) - \ell(h_{\mathbf{w}^*}, z_i) \leq \rho\|\mathbf{w}^i - \mathbf{w}^*\| \quad \text{und} \quad \ell(h_{\mathbf{w}^i}, z') - \ell(h_{\mathbf{w}^*}, z') \leq \rho\|\mathbf{w}^i - \mathbf{w}^*\| \ .$$

Also

$$2\rho\|\mathbf{w}^i - \mathbf{w}^*\| \geq m\sigma\|\mathbf{w}^i - \mathbf{w}^*\|^2 \ ,$$

[2] Ein technischer Unterschied ist, ob die (nun versteckte) Verschiebung der Hyperebene auch regularisiert wird oder nicht. Wir ignorieren dies.

und somit

$$\|\mathbf{w}^i - \mathbf{w}^*\| \leq \frac{2\rho}{m\sigma} \ .$$

Das heißt, es gilt auch

$$\ell(h_{S^i}, z_i) - \ell(h_S, z_i) = \ell(h_{\mathbf{w}^i}, z_i) - \ell(h_{\mathbf{w}^*}, z_i) \leq \rho\|\mathbf{w}^i - \mathbf{w}^*\| \leq \frac{2\rho^2}{m\sigma} \ .$$

□

4 Fazit

Wie wir gesehen haben, kann Regularisierung also Overfitting vermeiden. Anzumerken ist jedoch, dass die Regularisierung nicht zu stark gewählt werden darf. Anderenfalls wird der Trainingsfehler groß, es tritt also Underfitting ein.

Referenzen

- Understanding Machine Learning, Kapitel 13.3–13.4
- Foundations of Machine Learning, Kapitel 14.3 (weitergehend)