Machine learning lecture slides

COMS 4771 Fall 2020

Regression II: Regularization

Outline

- Inductive biases in linear regression
- Regularization
- Model averaging
- Bayesian interpretation of regularization

Inductive bias

- In linear regression, possible for least square solution to be non-unique, in which case there are infinitely-many solutions.
- Which one should we pick?
  - Possible answer: Pick shortest solution, i.e., of minimum (squared) Euclidean norm $\|w\|_2^2$.
  - Small norm $\Rightarrow$ small changes in output in response to changes in input:

$$\underbrace{|w^\intercal x - w^\intercal x'|}_{\text{change in output}} \leq \|w\|_2 \cdot \underbrace{\|x - x'\|_2}_{\text{change in input}}$$

    (easy consequence of Cauchy-Schwarz)
  - Note: data does not give reason to choose shorter w over longer w.
  - Preference for short w is an example of an *inductive bias*.
- All learning algorithms encode some form of inductive bias.

Example of minimum norm inductive bias

- Trigonometric feature expansion

$$\varphi(x) = (\sin(x), \cos(x), \ldots, \sin(32x), \cos(32x)) \in \mathbb{R}^{64}$$

  - $n = 32$ training examples
  - Infinitely many solutions to normal equations

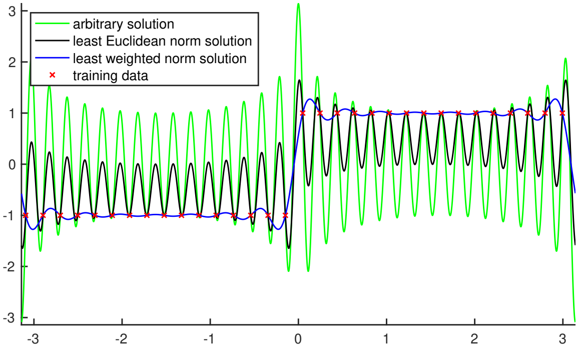


Figure 1: Fitted linear models with trigonometric feature expansion

Representation of minimum norm solution (1)

- **Claim**: The minimum (Euclidean) norm solution to normal equations lives in span of the x_i's (i.e., in $\mathrm{range}(A^\mathsf{T})$).
  - I.e., can write
    $$w = A^\mathsf{T}\alpha = \sum_{i=1}^{n} \alpha_i x_i$$
    for some $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$.
  - (Replace x_i with $\varphi(x_i)$ if using feature map φ.)

- **Proof**: If we have any solution of the form $w = s + r$, where $s \in \mathrm{range}(A^\mathsf{T})$, and $r \neq 0$ is in $\mathrm{null}(A)$ (i.e., $Ar = 0$), we can remove r and have a shorter solution:
  $$A^\mathsf{T}b = A^\mathsf{T}Aw = A^\mathsf{T}A(s + r) = A^\mathsf{T}As + A^\mathsf{T}(Ar) = A^\mathsf{T}As.$$
  (Recall Pythagorean theorem: $\|w\|_2^2 = \|s\|_2^2 + \|r\|_2^2$)

Representation of minimum norm solution (2)

- In fact, minimum Euclidean norm solution is unique!
  - If two distinct solutions w and w' have the same length, then averaging them gives another solution $\frac{1}{2}(w + w')$ of shorter length.

Regularization

- Combine two concerns: making both $\widehat{\mathcal{R}}(w)$ and $\|w\|_2^2$ small
  - Pick $\lambda \geq 0$, and minimize

$$\widehat{\mathcal{R}}(w) + \lambda\|w\|_2^2$$

- If $\lambda > 0$, solution is always unique (even if $n < d$).
  - Called *ridge regression*.
  - $\lambda = 0$ is OLS/ERM.
  - λ controls how much to pay attention to *regularizer* $\|w\|_2^2$ relative to *data fitting term* $\widehat{\mathcal{R}}(w)$
  - λ is hyperparameter to tune (e.g., using cross-validation)

- Solution is also in span of the x_i's (i.e., in $\mathrm{range}(A^\mathsf{T})$)

Example of regularization with squared norm penality

- Trigonometric feature expansion

$$\varphi(x) = (\sin(x), \cos(x), \ldots, \sin(32x), \cos(32x))$$

- Trade-off between fit to data and regularizer

$$\min_{w \in \mathbb{R}^{64}} \frac{1}{n} \sum_{i=1}^{n} \left(w^{\intercal} \varphi(x_i) - y_i\right)^2 + \lambda \sum_{j=1}^{32} 2^j (w_{\sin,j}^2 + w_{\cos,j}^2)$$

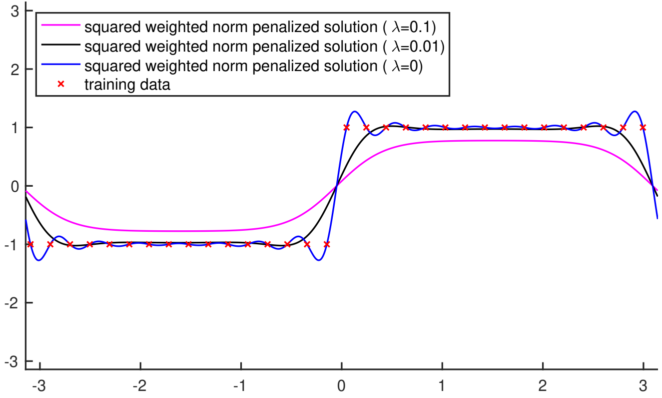


Figure 2: Trading-off between data fitting term and regularizer

Data augmentation (1)

- Let $\widetilde{A} = \begin{bmatrix} A \\ \sqrt{\lambda} I \end{bmatrix} \in \mathbb{R}^{(n+d)\times d}$ and $\tilde{b} = \begin{bmatrix} b \\ 0 \end{bmatrix} \in \mathbb{R}^{n+d}$
- Then $\|\widetilde{A}w - \tilde{b}\|_2^2 = \widehat{\mathcal{R}}(w) + \lambda\|w\|_2^2$ (ridge regression objective)

- Interpretation:
  - d "fake" data points, ensures augmented $\widetilde{A}$ has rank d
  - All corresponding labels are zero.
- $\widetilde{A}^\mathsf{T}\widetilde{A} = A^\mathsf{T}A + \lambda I$ and $\widetilde{A}^\mathsf{T}\tilde{b} = A^\mathsf{T}b$
- So ridge regression solution is $\hat{w} = (A^\mathsf{T}A + \lambda I)^{-1}A^\mathsf{T}b$

Data augmentation (2)

- Domain-specific data augmentation: e.g., image transformations

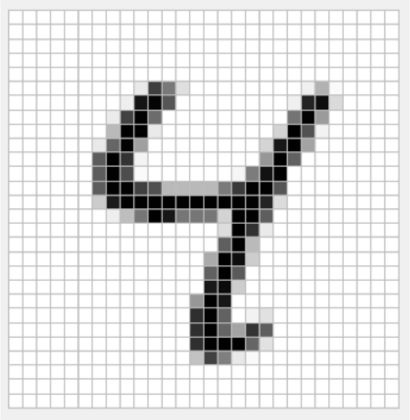


Figure 3: What data augmentations make sense for OCR digit recognition?

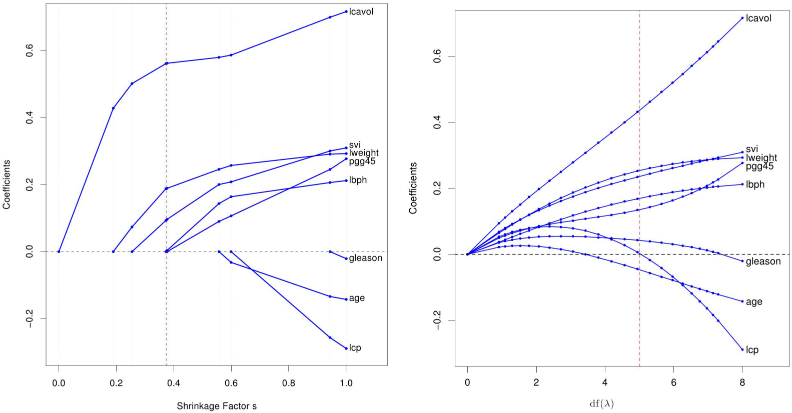
Lasso

- *Lasso*: minimize $\widehat{\mathcal{R}}(w) + \lambda \|w\|_1$
  - Here, $\|v\|_1 = \sum_{j=1}^{n} |v_j|$, sum of absolute values of vector entries
  - Prefers short w, where length is measured using different norm
  - Tends to produce w that are *sparse* (i.e., have few non-zero entries), or at least are well-approximated by sparse vectors.
  - A different inductive bias:

$$|w^{\mathsf{T}} x - w^{\mathsf{T}} x'| \leq \|w\|_1 \cdot \|x - x'\|_\infty$$

Lasso vs ridge regression

- Example: coefficient profile of Lasso vs ridge
- x = clinical measurements, y = level of prostate cancer antigen
- Horizontal axis: varying λ (large λ to left, small λ to right).
- Vertical axis: coefficient value in Lasso and ridge solutions, for eight different features



Inductive bias from minimum ℓ_1 norm

- **Theorem**: Pick any $w \in \mathbb{R}^d$ and any $\varepsilon \in (0,1)$. Form $\tilde{w} \in \mathbb{R}^d$ by including the $\lceil 1/\varepsilon^2 \rceil$ largest (by magnitude) coefficients of w, and setting remaining entries to zero. Then

$$\|\tilde{w} - w\|_2 \leq \varepsilon \|w\|_1.$$

  - If $\|w\|_1$ is small (compared to $\|w\|_2$), then theorem says w is well-approximated by sparse vector.

Sparsity

- Lasso also tries to make coefficients small. What if we only care about sparsity?
- *Subset selection*: minimize empirical risk among all k-sparse solutions
- Greedy algorithms: repeatedly choose new variables to "include" in support of w until k variables are included.
  - *Forward stepwise regression* / *orthogonal matching pursuit*: Each time you "include" a new variable, re-fit all coefficients for included variables.
  - Often works as well as Lasso
- Why do we care about sparsity?

Detour: Model averaging

- Suppose we have M real-valued predictors, $\hat{f}_1, \dots, \hat{f}_M$
- How to take advantage of all of them?
- *Model selection*: pick the best one, e.g., using hold-out method
- *Model averaging*: form "ensemble" predictor $\hat{f}_{\text{avg}}$, where for any x,

$$\hat{f}_{\text{avg}}(x) := \frac{1}{M} \sum_{j=1}^{M} \hat{f}_j(x).$$

Risk of model averaging

- $\mathcal{R}(f) := \mathbb{E}[(f(X) - Y)^2]$ for some random variable (X, Y) taking values in $\mathcal{X} \times \mathbb{R}$.
- **Theorem**: For any $\hat{f}_1, \ldots, \hat{f}_M \colon \mathcal{X} \to \mathbb{R}$, the ensemble predictor $\hat{f}_{\text{avg}} := \frac{1}{M} \sum_{j=1}^{M} \hat{f}_j$ satisfies

$$\mathcal{R}(\hat{f}_{\text{avg}}) = \frac{1}{M} \sum_{j=1}^{M} \mathcal{R}(\hat{f}_j) - \frac{1}{M} \sum_{j=1}^{M} \mathbb{E}\left[(\hat{f}_{\text{avg}}(X) - \hat{f}_j(X))^2\right].$$

- Better than model selection when:
  - all $\hat{f}_j$ have similar risks, and
  - all $\hat{f}_j$ predict very differently from each other

Stacking and features

- In model averaging, "weights" of $1/M$ for all $\hat{f}_j$ seems arbitrary
- Can "learn" weights using linear regression!
  - Use feature expansion $\varphi(x) = (\hat{f}_1(x), \ldots, \hat{f}_M(x))$
  - Called *stacking*
  - Use additional data (independent of $\hat{f}_1, \ldots, \hat{f}_M$)

- Upshot: Any function (even learned functions) can be a feature
- Conversely: Behind every feature is a deliberate modeling choice

Detour: Bayesian statistics

- *Bayesian inference*: probabilistic approach to updating beliefs
  - Posit a (parametric) statistical model for data (*likelihood*)
  - Start with some beliefs about the parameters of model (*prior*)
  - Update beliefs after seeing data (*posterior*)

$$\underbrace{\Pr(w \mid \mathsf{data})}_{\text{posterior}(w)} = \frac{1}{Z_{\mathsf{data}}} \underbrace{\Pr(w)}_{\text{prior}(w)} \cdot \underbrace{\Pr(\mathsf{data} \mid w)}_{\text{likelihood}(w)}$$

  - (Finding normalization constant Z_{data} is often the computationally challenging part of belief updating.)
  - Basis for reasoning in humans (maybe?), robots, etc.

Beyond Bayesian inference

- Can use Bayesian inference framework for designing estimation/learning algorithms (even if you aren't a Bayesian!)
  - E.g., instead of computing entire posterior distribution, find the w with highest posterior probability
  - Called *maximum a posteriori (MAP)* estimator
  - Just find w to maximize

$$\text{prior}(w) \times \text{likelihood}(w).$$

  - (Avoids issue with finding normalization constant.)

Bayesian approach to linear regression

- In linear regression model, express prior belief about $w = (w_1, \ldots, w_d)$ using a probability distribution with density function
  - Simple choice: $\text{prior}(w_1, \ldots, w_d) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{w_j^2}{2\sigma^2})$
  - I.e., treat $w_1, \ldots, w_d$ as independent $\text{N}(0, \sigma^2)$ random variables
- Likelihood model: $(X_1, Y_1), \ldots, (X_n, Y_n)$ are conditionally independent given w, and $Y_i \mid (X_i, w) \sim \text{N}(X_i^\mathsf{T} w, 1)$.
- What is the MAP?

MAP for Bayesian linear regression

- Find w to maximize

$$\underbrace{\prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{w_j^2}{2\sigma^2}\right)}_{\text{prior}(w)} \cdot \underbrace{\prod_{i=1}^{n} p(x_i) \cdot \frac{1}{\sqrt{2\pi}} \exp(-(y_i - x_i^{\mathsf{T}} w)^2/2)}_{\text{likelihood}(w)}.$$

  (Here, p is marginal density of X; unimportant.)
- Take logarithm and omit terms not involving w:

$$-\frac{1}{2\sigma^2} \sum_{i=1}^{d} w_j^2 - \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i^{\mathsf{T}} w)^2.$$

- For $\sigma^2 = \frac{1}{n\lambda}$, same as minimizing

$$\frac{1}{n} \sum_{i=1}^{n} (x_i^{\mathsf{T}} w - y_i)^2 + \lambda \|w\|_2^2,$$

  which is the ridge regression objective!

Example: Dartmouth data example

- Dartmouth data example, where we considered intervals for the HS GPA variable:

$$(0.00, 0.25], \quad (0.25, 0.50], \quad (0.50, 0.75], \quad \cdots$$

- Use $\varphi(x) = (\mathbf{1}_{\{x \in (0.00, 0.25]\}}, \mathbf{1}_{\{x \in (0.25, 0.50]\}}, \dots)$ with a linear function
- Regularization: $\lambda \sum_{j=1}^{d} (w_j - \mu)^2$ where $\mu = 2.46$ is mean of College GPA values.
- What's the Bayesian interpretation of minimizing the following objective?

$$\frac{1}{n} \sum_{i=1}^{n} (\varphi(x_i)^{\mathsf{T}} w - y_i)^2 + \lambda \sum_{j=1}^{d} (w_j - \mu)^2$$