Machine learning lecture slides

Regression II: Regularization

Outline

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

- 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
    2

change in output

(easy consequence of Cauchy-Schwa

  - Note: data does not give reason to choose sh 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

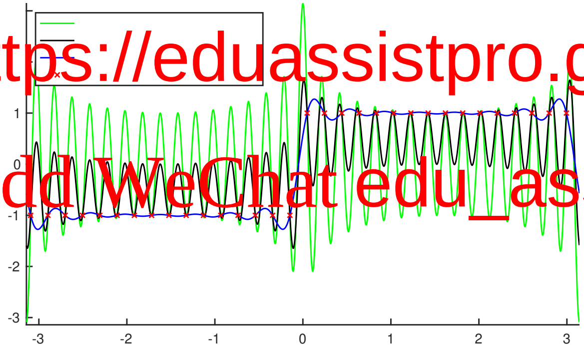

Figure 1: Fitted linear models with trigonometric feature expansion

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 $\mathrm{range}(A^\top)$).
  - I.e., can write

$$w = A^\top \alpha = \sum_{i=1}^{n} \alpha_i x_i$$

- **Proof**: If we have any solution of the form e $s \in \mathrm{range}(A^\top)$, and $r \neq 0$ is in n remove r and have a shorter solution:

$$A^\top b = A^\top A w = A^\top A(s + r) = A^\top A s + A^\top (Ar) = A^\top A s.$$

(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

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

-
  - $\lambda = 0$ is OLS/ERM.
  - λ controls how much to pay attention to
    relative to *data fitting term* $\widehat{\mathcal{R}}($

  - λ is hyperparameter to tune (e.g., using

- 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

[illegible]

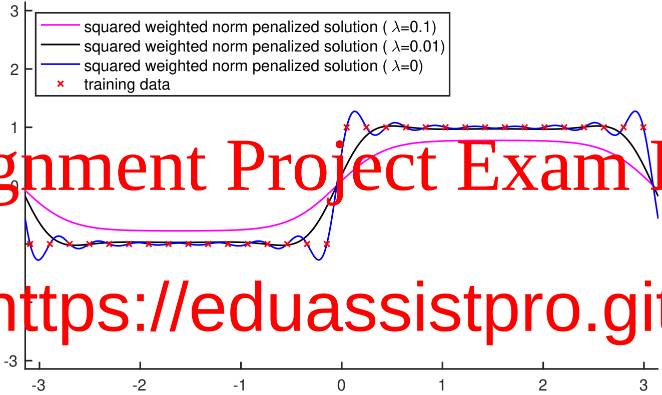

Figure 2: Trading-off between data fitting ter

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)

-
  - d "fake" data points, ensures augmen
  - All corresponding labels are zero

- $\widetilde{A}^\intercal \widetilde{A} = A^\intercal A + \lambda I$ and $\widetilde{A}^\intercal \tilde{b} = A^\intercal$
- So ridge regression solution is $\hat{w} = (A^\intercal A + \lambda I)^{-1} A^\intercal b$

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

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

- *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

∞

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

- **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

$$\tilde{w} \quad w \qquad \varepsilon \; w \quad .$$

$\tilde{w}$ is

- Lasso also tries to make coefficients small. What if we only care about sparsity?
- *Subset selection*: minimize empirical risk among all k-sparse solutions
-

  t:

  included variables.
  - Often works as well as Lasso
- Why do we care about sparsity?

- Suppose we have M real-valued predictors, $\hat{f}_1, \ldots, \hat{f}_M$
- 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 f_{avg}, where for

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 $\hat{f}_1, \ldots, \hat{f}_M : \mathcal{X} \to \mathbb{R}$, the ensemble predictor $\hat{f}_{\text{avg}} := \frac{1}{M} \sum_{j=1}^{M} \hat{f}_j$ satisfies

$$\ldots))^2]$$

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

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*

- Upshot: Any function (even learned functi
- Conversely: Behind every feature is a delibe
  choice

- *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*)

  - (Finding normalization constant
    computationally challenging part of b
  - Basis for reasoning in humans (maybe?

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

  - (Avoids issue with finding normalizati

- In linear regression model, express prior belief about $w = (w_1, \ldots, w_d)$ using a probability distribution with density function
  - Simple choice: $\mathrm{prior}(w_1, \ldots, w_d) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{w_j^2}{2\sigma^2})$

    ariables

    ally

-
-

- Find w to maximize

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

-

$$-\frac{1}{2\sigma^2}\sum_{j=1}^{d} w_j^2 - \frac{1}{2} \ldots_i$$

- 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!

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

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

- ear

- an of College GPA values.
- What's the Bayesian interpretation of mi objective?

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