

Regression—linear and otherwise

Hal Daumé III

CS5350: Machine Learning

12 February 2008

1. Linear regression
 - 1.1 The “optimal” model
 - 1.2 The regularized model
 - 1.3 Algebraic solution
 - 1.4 Non-algebraic solution
2. Support-vector regression

Linear regression—the task

Just like classification, except outputs are $y \in \mathbb{R}$.

1. In the **linear** case, want function: $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$
2. Function should satisfy: $f(\mathbf{x}_n) = y_n$ for all training points $(\mathbf{x}_n, y_n)_{n=1}^N$.
3. Why is this not a reasonable requirement?
4. Want to measure **loss** between predicted value and truth, $\ell(y, \hat{y})$

Regularized linear regression, square loss

Typical choice of loss function:

$$\ell(y, \hat{y}) = (y - \hat{y})^2$$

Called *square(d) loss*. Problem becomes:

$$\min_{\mathbf{w}, b} \frac{1}{2} \sum_n (\mathbf{w}^\top \mathbf{x}_n + b - y_n)^2$$

Helps to **regularize** the weights:

$$\min_{\mathbf{w}, b} \underbrace{\frac{1}{2} \sum_n (\mathbf{w}^\top \mathbf{x}_n + b - y_n)^2}_{\text{Loss}} + \frac{\lambda}{2} \underbrace{\|\mathbf{w}\|^2}_{\text{regularization}}$$

Solving regularized linear regression

For simplicity, consider *unbiased* case.

Matrixifying the problem, \mathbf{w} is $D \times 1$, \mathbf{x}_n is $D \times 1$, everything else is scalar. Let \mathbf{X} be the $N \times D$ matrix with $[\mathbf{X}]_{n,d} = x_{n,d}$, \mathbf{Y} be the $N \times 1$ vector of y_n s. Then:

$$\begin{aligned}
J(\mathbf{w}) &= \frac{1}{2} \sum_n (\mathbf{w}^\top \mathbf{x}_n - y_n)^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \\
&= \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{Y}\|^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2 \\
&= \frac{1}{2} [\mathbf{X}\mathbf{w} - \mathbf{Y}]^\top [\mathbf{X}\mathbf{w} - \mathbf{Y}] + \frac{\lambda}{2} \mathbf{w}^\top \mathbf{w} \\
&= \frac{1}{2} [\mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w} + \mathbf{Y}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X} \mathbf{w} - \mathbf{w}^\top \mathbf{X}^\top \mathbf{Y}] + \frac{\lambda}{2} \mathbf{w}^\top \mathbf{w} \\
&= \frac{1}{2} [\mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w} + \mathbf{Y}^\top \mathbf{Y} - 2\mathbf{Y}^\top \mathbf{X} \mathbf{w}] + \frac{\lambda}{2} \mathbf{w}^\top \mathbf{w}
\end{aligned}$$

CS5350

Hal Daumé III (U Utah)

5 / 12

Solving regularized linear regression

Computing derivatives ($\mathbf{w} : 1 \times D$, $\mathbf{X} : N \times D$, $\mathbf{Y} : N \times 1$)

$$\begin{aligned}
J(\mathbf{w}) &= \frac{1}{2} [\mathbf{w}^\top \mathbf{X}^\top \mathbf{X} \mathbf{w} + \mathbf{Y}^\top \mathbf{Y} - 2\mathbf{Y}^\top \mathbf{X} \mathbf{w}] + \frac{\lambda}{2} \mathbf{w}^\top \mathbf{w} \\
\nabla_{\mathbf{w}} J &= \frac{1}{2} [2\mathbf{X}^\top \mathbf{X} \mathbf{w} - 2\mathbf{X}^\top \mathbf{Y}] + \lambda \mathbf{w} \\
&= \mathbf{X}^\top \mathbf{X} \mathbf{w} - \mathbf{X}^\top \mathbf{Y} + \lambda \mathbf{w} = 0 \\
&\Rightarrow \mathbf{X}^\top \mathbf{X} \mathbf{w} + \lambda \mathbf{w} = \mathbf{X}^\top \mathbf{Y} \\
&\Rightarrow [\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}] \mathbf{w} = \mathbf{X}^\top \mathbf{Y} \\
&\Rightarrow \mathbf{w} = [\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}]^{-1} \mathbf{X}^\top \mathbf{Y}
\end{aligned}$$

A totally closed form solution!

CS5350

Hal Daumé III (U Utah)

6 / 12

The gradient descent solution

We have the regularized loss function:

$$\min_{\mathbf{w}, b} \sum_n (\mathbf{w}^\top \mathbf{x}_n + b - y_n)^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

We can easily take derivatives:

$$\begin{aligned}
\nabla_{\mathbf{w}} &= \sum_n 2(\mathbf{w}^\top \mathbf{x}_n + b - y_n) \mathbf{x}_n + \lambda \mathbf{w} \\
\nabla_b &= \sum_n 2(\mathbf{w}^\top \mathbf{x}_n + b - y_n)
\end{aligned}$$

while not converged:

- Update: $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla_{\mathbf{w}}$
- Update: $b \leftarrow b - \eta \nabla_b$

CS5350

Hal Daumé III (U Utah)

7 / 12

Bias/variance tradeoff in linear regression

Let y_n be the observed point, t_n be the truth (so $y_n = t_n + \text{noise}$) and let \hat{y}_n be our prediction. Call ϵ the noise. Then our expected loss is:

$$\frac{1}{2} \sum_n \mathbb{E} [(t_n - \hat{y}_n)^2]$$

A bit of algebra yields:

$$\begin{aligned}
\mathbb{E} [(t_n - \hat{y}_n)^2] &= \mathbb{E} [\epsilon^2] + \mathbb{E} [(t_n - \mathbb{E}[\hat{y}_n])^2] + \mathbb{E} [(\mathbb{E}[\hat{y}_n] - \hat{y}_n)^2] \\
&= \mathbb{V}[\text{noise}] + \text{bias}^2 + \mathbb{V}[\hat{y}_n]
\end{aligned}$$

So our error is the sum of the variance of the noise (we can't control), the (squared) bias of our predictor and its own variance.

\Rightarrow Bias/variance trade-off

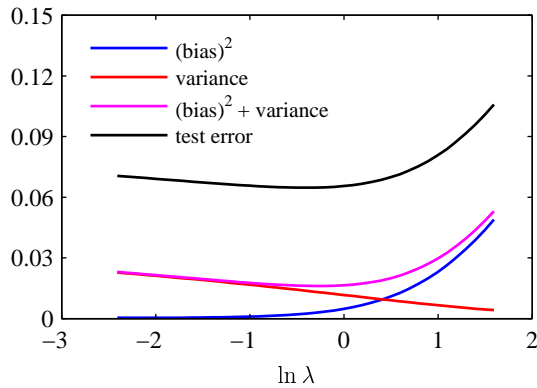
CS5350

Hal Daumé III (U Utah)

8 / 12

Bias/variance trade-off in linear regression

$$\begin{aligned}\mathbb{E}[(t_n - f_n)^2] &= \mathbb{E}[\epsilon^2] + \mathbb{E}[(t_n - \mathbb{E}[\hat{y}_n])^2] + \mathbb{E}[(\mathbb{E}[\hat{y}_n] - \hat{y}_n)^2] \\ &= \mathbb{V}[\text{noise}] + \text{bias}^2 + \mathbb{V}[\hat{y}_n]\end{aligned}$$



CS5350

Hal Daumé III (U Utah)

9 / 12

Support vector regression

Idea: extend hinge loss to regression.

Support vector classification:

$$\begin{aligned}\text{minimize}_{\mathbf{w}} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{subject to} \quad & y_n [\mathbf{w}^\top \mathbf{x}_n + b] - 1 \geq 0 \quad (\forall n)\end{aligned}$$

Support vector regression (ϵ -insensitive loss):

$$\begin{aligned}\text{minimize}_{\mathbf{w}} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{subject to} \quad & \mathbf{w}^\top \mathbf{x}_n + b \geq y_n - \epsilon \quad (\forall n) \\ & \mathbf{w}^\top \mathbf{x}_n + b \leq y_n + \epsilon \quad (\forall n)\end{aligned}$$

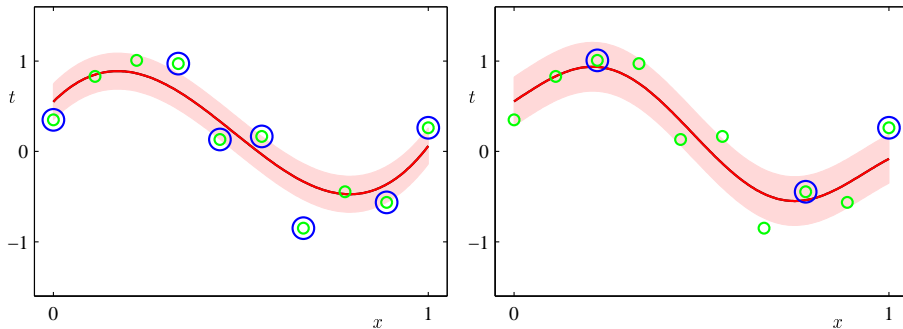
Can be kernelized...

CS5350

Hal Daumé III (U Utah)

10 / 12

ϵ -insensitive regression



CS5350

Hal Daumé III (U Utah)

11 / 12

Support vector regression

Hard margin version:

$$\begin{aligned}\text{minimize}_{\mathbf{w}} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{subject to} \quad & \mathbf{w}^\top \mathbf{x}_n + b \geq y_n - \epsilon \quad (\forall n) \\ & \mathbf{w}^\top \mathbf{x}_n + b \leq y_n + \epsilon \quad (\forall n)\end{aligned}$$

Slack-ified version:

$$\begin{aligned}\text{minimize}_{\mathbf{w}} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_n \xi_n \\ \text{subject to} \quad & \mathbf{w}^\top \mathbf{x}_n + b \geq y_n - \epsilon - \xi_n \quad (\forall n) \\ & \mathbf{w}^\top \mathbf{x}_n + b \leq y_n + \epsilon + \xi_n \quad (\forall n) \\ & \xi_n \geq 0 \quad (\forall n)\end{aligned}$$

CS5350

Hal Daumé III (U Utah)

12 / 12