Pattern Recognition 2015
Linear Models Selection and Regularization

Ad Feelders

Universiteit Utrecht

November 25, 2015

Notation

Bishop/Feelders	James	Meaning
t	Y	target/dependent variable
y	$\hat{y}$	fitted value/prediction
$\mathbf{w}$	β	weights/coefficients
$\mathbf{w}_{\mathrm{ML}}$, $\mathbf{w}$ (sloppy!)	$\hat{\beta}$	coefficient estimates
N	n	number of observations
SSE	RSS	sum of squared errors/residual sum of squares
SST	TSS	total sum of squares

Bias-Variance Decomposition

To understand some of the material of Chapter 6 of James et al., it is useful to know about the bias-variance decomposition of prediction error in regression.

Study section 3.2 of Bishop to this end. The next slides may be useful in studying this part of the book of Bishop. Equation numbers on the slides refer to equations in the book of Bishop.

Expectation and Variance

Some elementary properties:

1. $\mathbb{E}[c] = c$ for constant c.
2. $\mathbb{E}[cx] = c\mathbb{E}[x]$.
3. $\mathbb{E}[x \pm y] = \mathbb{E}[x] \pm \mathbb{E}[y]$.
4. $\text{var}[c] = 0$.
5. $\text{var}[cx] = c^2\text{var}[x]$.
6. $\text{var}[x \pm y] = \text{var}[x] + \text{var}[y]$ if x and y independent.

Bias of an Estimator

If $\hat{\theta}$ denotes an estimator of θ, then the estimation error is a random variable $\hat{\theta} - \theta$, which should preferably be close to zero.
Bias of $\hat{\theta}$:

$$B[\hat{\theta}] = \mathbb{E}[\hat{\theta} - \theta] = \mathbb{E}[\hat{\theta}] - \theta,$$

where expectation is taken with respect to repeated samples.

If $\mathbb{E}[\hat{\theta}] = \theta$, the estimator $\hat{\theta}$ is called *unbiased*.

Variance of an Estimator

Another important quality measure is variance:

$$\text{var}[\hat{\theta}] = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$$

which measures how much individual estimates tend to differ from $\mathbb{E}[\hat{\theta}]$.

Example: Bernoulli Distribution

Suppose $p(x = 1) = \mu$ where $x \in \{0, 1\}$, or

$$p(x) = \mu^x (1 - \mu)^{(1-x)} \tag{2.2}$$

Note that

$$\mathbb{E}[x] = \sum_x x p(x) = 0 \cdot (1 - \mu) + 1 \cdot \mu = \mu \tag{2.3}$$

and

$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \mu - \mu^2 = \mu(1 - \mu) \tag{2.4}$$

Example: Bernoulli Distribution

Random sample of size N from this population. Show that

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{2.7}$$

is an unbiased estimator of μ.

Note: on each draw x_n has expected value $\mathbb{E}[x_n] = \mu$ and $\mathrm{var}[x_n] = \mu(1-\mu)$.

Example: Bernoulli Distribution

μ_{ML} is unbiased

$$\mathbb{E}[\mu_{\text{ML}}] = \mathbb{E}\left[\frac{1}{N}\sum x_n\right] = \frac{1}{N}\sum \mathbb{E}[x_n] = \frac{1}{N}N\mu = \mu$$

Variance of μ_{ML}

$$\begin{aligned} \text{var}[\mu_{\text{ML}}] &= \text{var}\left[\frac{1}{N}\sum x_n\right] = \frac{1}{N^2}\sum \text{var}[x_n] \\ &= \frac{1}{N^2}N\mu(1-\mu) = \frac{\mu(1-\mu)}{N} \end{aligned}$$

Mean Square Error and its Decomposition

Overall quality measure:

$$M[\hat{\theta}] = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

where low values indicate a good estimator.

We can decompose mean squared error into

$$M[\hat{\theta}] = B[\hat{\theta}]^2 + \mathrm{var}[\hat{\theta}]$$

Proof of Decomposition

Write $\hat{\theta} - \theta$ as

$$\hat{\theta} - \theta = (\hat{\theta} - \mathbb{E}[\hat{\theta}]) + (\mathbb{E}[\hat{\theta}] - \theta)$$

Square on left and right

$$\begin{aligned}(\hat{\theta} - \theta)^2 &= (\hat{\theta} - \mathbb{E}[\hat{\theta}])^2 + (\mathbb{E}[\hat{\theta}] - \theta)^2 \\ &+ 2(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta)\end{aligned}$$

Proof of Decomposition (continued)

Take expectations left and right.

Since

$$\mathbb{E}[\hat{\theta} - \mathbb{E}[\hat{\theta}]] = \mathbb{E}[\hat{\theta}] - \mathbb{E}[\hat{\theta}] = 0$$

and $\mathbb{E}[\hat{\theta}] - \theta$ is a constant, the cross term drops out.
So we get

$$\begin{aligned} \mathbb{E}[(\hat{\theta} - \theta)^2] &= \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2] + (\mathbb{E}[\hat{\theta}] - \theta)^2 \\ &= \mathrm{var}[\hat{\theta}] + B[\hat{\theta}]^2 \end{aligned}$$

Mean Square Prediction Error

For fixed $\mathcal{D}$ and $\mathbf{x}$ the mean square prediction error of $y(\mathbf{x}, \mathcal{D})$ is given by:

$$\mathbb{E}[(t - y(\mathbf{x}; \mathcal{D}))^2]$$

where the expectation is taken with respect to $p(t|\mathbf{x})$.

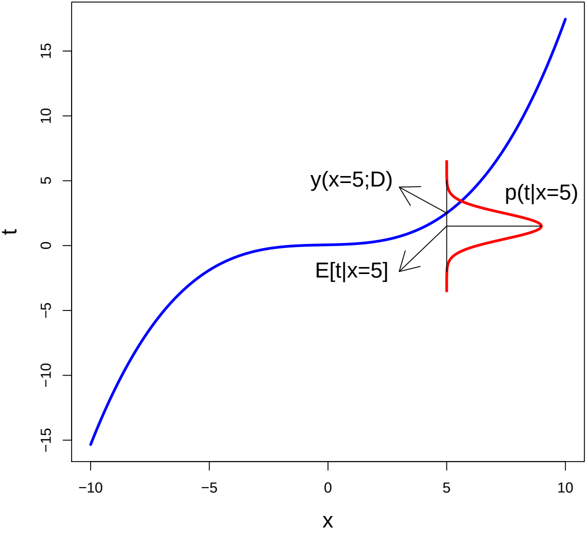
Decomposition of MSPE

Recall that $h(\mathbf{x}) \equiv \mathbb{E}[t|\mathbf{x}]$ is the best possible predictor of t at $\mathbf{x}$.

Decomposition of MSPE of $y(\mathbf{x}; \mathcal{D})$:

$$\mathbb{E}[(t - y(\mathbf{x}; \mathcal{D}))^2] = \underbrace{(h(\mathbf{x}) - y(\mathbf{x}; \mathcal{D}))^2}_{\text{reducible}} + \underbrace{\mathbb{E}[(t - h(\mathbf{x}))^2]}_{\text{irreducible}}$$

Decomposition of MSPE



Proof: expand left-hand side

Write h for $h(\mathbf{x})$ and y for $y(\mathbf{x};\mathcal{D})$.

$$\begin{aligned}\mathbb{E}[(t-y)^2] &= \mathbb{E}[t^2 - 2ty + y^2] \\ &= \mathbb{E}[t^2|\mathbf{x}] - 2y\mathbb{E}[t|\mathbf{x}] + y^2 \\ &= \mathbb{E}[t^2|\mathbf{x}] - 2yh + y^2\end{aligned}$$

since $\mathbb{E}[t|\mathbf{x}] \equiv h$

Proof: expand right-hand side

$$\begin{aligned}(h-y)^2 &+& \mathbb{E}[(t-h)^2] = h^2 - 2hy + y^2 + \mathbb{E}[t^2|\mathbf{x}] \\ &-& 2h\mathbb{E}[t|\mathbf{x}] + h^2 = \mathbb{E}[t^2|\mathbf{x}] - 2yh + y^2\end{aligned}$$

since $\mathbb{E}[t|\mathbf{x}] \equiv h$

Decomposition of Reducible Part

Let $h \equiv h(\mathbf{x})$ and $y \equiv y(\mathbf{x}; \mathcal{D})$. The mean square estimation error of y as estimator of h is given by

$$\mathbb{E}_{\mathcal{D}}[(h - y)^2] = (h - \mathbb{E}_{\mathcal{D}}[y])^2 + \mathbb{E}_{\mathcal{D}}[(y - \mathbb{E}_{\mathcal{D}}[y])^2] \qquad (3.40)$$

Or, in other "words":

$$M[y] = B[y]^2 + \mathrm{var}[y],$$

Mean square error = squared bias + variance