EN.600.475 Machine Learning

Linear Regression

Raman Arora
Lecture 6
February 15, 2017

- Bayes predictor, error decomposition
- Gaussian noise model

Slides credit: Greg Shakhnarovich

Review: loss and risk

- Assume that data are sampled from (unknown) $p(\mathbf{x}, y)$
- Choose loss function L, parametric model family $f(\mathbf{x}; \mathbf{w})$
- The ultimate goal is to minimize the *expected loss*, also known as *risk*:

$$R(\mathbf{w}) \;=\; E_{(\mathbf{x}_0, y_0) \sim p(\mathbf{x}, y)} \left[\ell \left(f(\mathbf{x}_0; \mathbf{w}), y_0 \right) \right]$$

- Measurable proxy: empirical loss on training set

$$\widehat{R}_n(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \ell \left(f(\mathbf{x}_i; \mathbf{w}), y_i \right)$$

- This is called empirical risk minimization (ERM)

Review: least squares linear regression

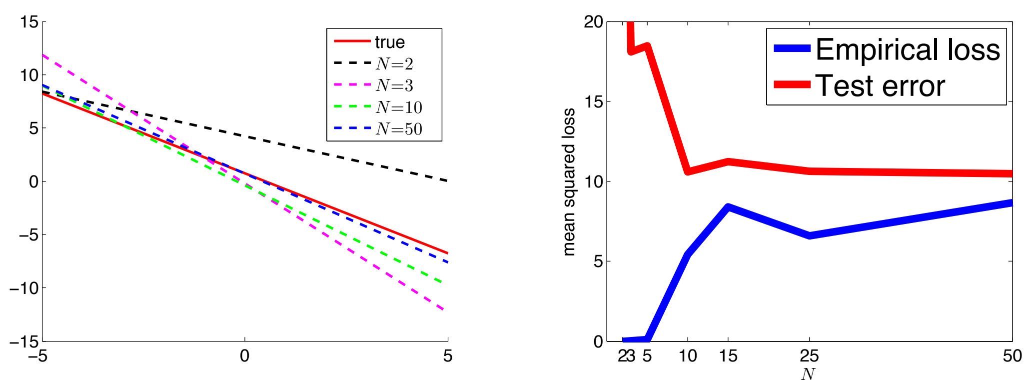
- Mapping $f:\ \mathbf{x} \in \mathbb{R}^d \to y \in \mathbb{R}$
- Two choices: linear model $f(\mathbf{x};\mathbf{w}) = \mathbf{w}^T\mathbf{x}$, and squared loss $\ell(\widehat{y}, y) = (\widehat{y} - y)^2$
- Least squares fitting:

$$\underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{N} (\mathbf{w}\cdot\mathbf{x}_i - y_i)^2$$
$$\Rightarrow \mathbf{w}^* = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

- Computationally: need to compute pseudo-inverse of the data matrix $\mathbf{X}$

Linear regression - generalization

- Toy experiment: fit a line to varying number of points drawn from the same distribution $p(\mathbf{x}, y)$



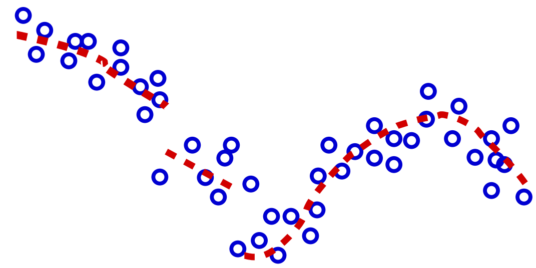
- A paradox?
  - The more training data we have, the "worse" is the fit;
  - But at the same time our prediction ability seems to improve.

Best unrestricted predictor

- What is the *best possible* predictor of y, in terms of expected squared loss, if we do not restrict $\mathcal{H}$ at all?

$$f^* = \operatorname*{argmin}_{f:\mathcal{X}\to\mathbb{R}} E_{(\mathbf{x}_0,y_0)\sim p(\mathbf{x},y)}\left[(f(\mathbf{x}_0)-y_0)^2\right]$$

- *Any* $f:\mathcal{X}\to\mathbb{R}$ is allowed.



The *chain rule* of probability: $p(\mathbf{x},y)=p(y|\mathbf{x})p(\mathbf{x})$

By definition: $E_{p(y,\mathbf{x})}\left[g(y,\mathbf{x})\right]=\int_{\mathbf{x}}\int_{y} g(y,\mathbf{x})p(y|\mathbf{x})p(\mathbf{x})dyd\mathbf{x}$

$$E_{(\mathbf{x}_0,y_0)\sim p(\mathbf{x},y)}\left[(f(\mathbf{x}_0)-y_0)^2\right] = E_{\mathbf{x}_0\sim p(\mathbf{x})}\left[E_{y_0\sim p(y|\mathbf{x})}\left[(f(\mathbf{x}_0)-y_0)^2 \mid \mathbf{x}_0\right]\right]$$

Best unrestricted predictor

$$\begin{aligned}E_{(\mathbf{x}_0,y_0)\sim p(\mathbf{x},y)}\left[(f(\mathbf{x}_0)-y_0)^2\right] &= E_{\mathbf{x}_0\sim p(\mathbf{x})}\left[E_{y_0\sim p(y|\mathbf{x})}\left[(f(\mathbf{x}_0)-y_0)^2 \mid \mathbf{x}_0\right]\right]\\ &= \int_{\mathbf{x}_0}\left\{E_{y_0\sim p(y|\mathbf{x})}\left[(f(\mathbf{x}_0)-y_0)^2 \mid \mathbf{x}_0\right]\right\}p(\mathbf{x}_0)d\mathbf{x}_0\end{aligned}$$

- Must minimize the inner conditional expectation for each $\mathbf{x}_0$!

$$\begin{aligned}\frac{\delta}{\delta f(\mathbf{x})}E_{p(y|\mathbf{x})}\left[(f(\mathbf{x}_0)-y_0)^2 \mid \mathbf{x}_0\right] &= 2E_{p(y|\mathbf{x})}\left[f(\mathbf{x}_0)-y_0 \mid \mathbf{x}_0\right]\\ &= 2\left(f(\mathbf{x}_0)-E_{p(y|\mathbf{x})}\left[y_0|\mathbf{x}_0\right]\right)=0\end{aligned}$$

- We minimize the expected loss by setting f to the conditional expectation of y for each $\mathbf{x}$:

$$f^*(\mathbf{x}_0) = E_{p(y|\mathbf{x}_0)}\left[y_0|\mathbf{x}_0\right]$$

Generative versus discriminative approach

- Conceptually, if we know $p\left(y \mid \mathbf{x}\right)$ we can find the best unrestricted predictor by taking for each $\mathbf{x}_0$ the expectation

$$\hat{y}(\mathbf{x}_0) \;=\; f(\mathbf{x}_0) \;=\; E_{y \sim p(y \mid \mathbf{x}_0)}\left[y|\mathbf{x}_0\right]$$

- Generative approach:
  - Estimate the joint probability density $p(\mathbf{x}, y)$
  - *Normalize* to find the conditional density $p(y|\mathbf{x})$
- Discriminative approach:
  - Estimate/infer the conditional density $p(y|\mathbf{x})$ *directly* from the data; don't bother with $p(\mathbf{x}, y)$.
- Non-probabilistic approach: don't deal with probabilities, fit $f(\mathbf{x})$ directly to the data.

Decomposition of error

Let's take a closer look at the expected loss:

- $\hat{\mathbf{w}}$ are LSQ estimates from training data.
- $\mathbf{w}^*$ are *optimal* linear regression parameters (generally unknown!)
- $y - \hat{\mathbf{w}} \cdot \mathbf{x} \;=\; (y - \mathbf{w}^* \cdot \mathbf{x}) + (\mathbf{w}^* \cdot \mathbf{x} \;-\; \hat{\mathbf{w}} \cdot \mathbf{x})$

$$\begin{aligned} E_{p(\mathbf{x},y)}\left[(y - \hat{\mathbf{w}} \cdot \mathbf{x})^2\right] \;=\; & E_{p(\mathbf{x},y)}\left[(y - \mathbf{w}^* \cdot \mathbf{x})^2\right] \\ & + \; 2E_{p(\mathbf{x},y)}\left[(y - \mathbf{w}^* \cdot \mathbf{x})\left(\mathbf{w}^* \cdot \mathbf{x} \;-\; \hat{\mathbf{w}} \cdot \mathbf{x}\right)\right] \\ & + \; E_{p(\mathbf{x},y)}\left[(\mathbf{w}^* \cdot \mathbf{x} \;-\; \hat{\mathbf{w}} \cdot \mathbf{x})^2\right]. \end{aligned}$$

- The second term vanishes since prediction errors $y - \mathbf{w}^* \cdot \mathbf{x}$ are uncorrelated with *any* linear function of $\mathbf{x}$ including $\mathbf{w}^* \cdot \mathbf{x} \;-\; \hat{\mathbf{w}} \cdot \mathbf{x}$.

Decomposition of error

$$E_{p(\mathbf{x},y)}\left[(y - \hat{\mathbf{w}} \cdot \mathbf{x})^2\right] = E_{p(\mathbf{x},y)}\left[(y - \mathbf{w}^* \cdot \mathbf{x})^2\right] + E_{p(\mathbf{x},y)}\left[(\mathbf{w}^* \cdot \mathbf{x} - \hat{\mathbf{w}} \cdot \mathbf{x})^2\right].$$

- *Approximation error* $E_{p(\mathbf{x},y)}\left[(y - \mathbf{w}^* \cdot \mathbf{x})^2\right]$ measures inherent limitations of the chosen hypothesis class (linear function). This error will remain even with infinite training data.
- *Estimation error* $E_{p(\mathbf{x},y)}\left[(\mathbf{w}^* \cdot \mathbf{x} - \hat{\mathbf{w}} \cdot \mathbf{x})^2\right]$ measures how close to the optimal $\mathbf{w}^*$ is $\hat{\mathbf{w}}$ estimated from (finite) training data.
- Note: since training data X, Y are random variables drawn from $p(\mathbf{x}, y)$, the estimated $\hat{\mathbf{w}}$ is a random variable as well.

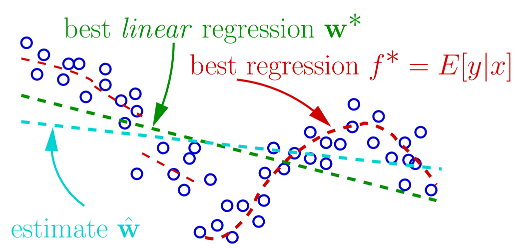
Decomposition of error

- Approximation error

$$E\left[(y - \mathbf{w}^* \cdot \mathbf{x})^2\right]$$

- Estimation error

$$E\left[(\mathbf{w}^* \cdot \mathbf{x} - \hat{\mathbf{w}} \cdot \mathbf{x})^2\right]$$



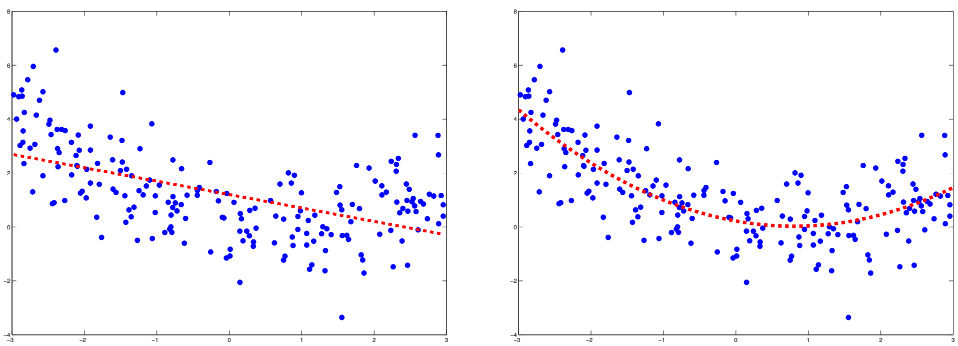
- For a *consistent* estimation procedure, $\lim_{N\to\infty} \hat{\mathbf{w}} = \mathbf{w}^*$, and so the estimation error decreases to zero with N.
- The approximation error can not be removed without changing the hypothesis class
- Approximation error depends on f^*. If $f^* \in \mathcal{H}$, it is minimized; is it zero then?

Statistical view of regression

- We will now explicitly model the randomness in the data:

$$y = f(\mathbf{x}; \mathbf{w}) + \nu$$

where the *noise* ν accounts for everything not captured by f.



- Quadratic component is noise on the left (linear model), part of signal on the right (quadratic model).

Statistical view of regression

$$y = f(\mathbf{x}; \mathbf{w}) + \nu$$

- Under this model, the best predictor is

$$E_{p(y|\mathbf{x})}\left[f(\mathbf{x}; \mathbf{w}) + \nu \,|\, \mathbf{x}\right] = f(\mathbf{x}; \mathbf{w}) + E_{p(\nu)}\left[\nu\right]$$

- Typically, $E_{p(\nu)}\left[\nu\right] = 0$ (*white* noise).
- Under such a model, $f(\mathbf{x}; \mathbf{w})$ captures the expected value of $y|\mathbf{x}$ *if* we believe the distribution in the model.
  - If the model is "correct", f is optimal.
  - Real data unlikely to have a "correct" parametric model.

Gaussian noise model

$$y = f(\mathbf{x}; \mathbf{w}) + \nu, \quad \nu \sim \mathcal{N}\left(\nu;\, 0, \sigma^2\right)$$

- Given the input $\mathbf{x}$, the label y is a random variable

  $$p(y|\mathbf{x}; \mathbf{w}, \sigma) \;=\; \mathcal{N}\left(y;\, f(\mathbf{x}; \mathbf{w}), \sigma^2\right)$$

  that is,

  $$p(y|\mathbf{x}; \mathbf{w}, \sigma) \;=\; \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y - f(\mathbf{x}; \mathbf{w}))^2}{2\sigma^2}\right)$$

- This is an explicit model of y that allows us, for instance, to *sample* y for a given $\mathbf{x}$.

Likelihood

- The *likelihood* of the parameters $\mathbf{w}$ given the observed data $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N], Y = [y_1, \ldots, y_N]^T$ is defined as

  $$p(Y|X; \mathbf{w}, \sigma)$$

  i.e., the probability of observing these ys for the given $\mathbf{x}$s, under the model parametrized by $\mathbf{w}$ and σ.
- Under the assumption that data are i.i.d. (independently, identically distributed) according to $p(\mathbf{x})$,

  $$p(Y|X; \mathbf{w}, \sigma) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i, \mathbf{w}, \sigma)$$

Maximum likelihood estimation

- *Maximum likelihood* (ML) estimation principle:

$$\hat{\mathbf{w}}_{ML} = \operatorname*{argmax}_{\mathbf{w}} p(Y|X;\mathbf{w},\sigma)$$

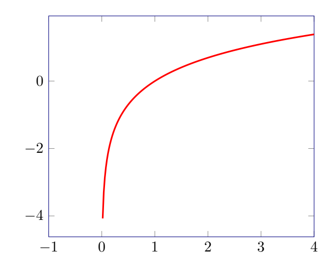
- Here we focus on likelihood as a function of $\mathbf{w}$.
- For Gaussian noise model:

$$\hat{\mathbf{w}}_{ML} = \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - f(\mathbf{x}_i;\mathbf{w}))^2}{2\sigma^2}\right)$$

- This may become numerically unwieldy...

Log-likelihood

- Properties of $\log$:
  - Defined for any $x > 0$.
  - Monotonically increasing.
  - $\log(AB) = \log A + \log B$,
    $\log A^B = B \log A$.



- Maximum likelihood $\max_{\mathbf{w}} p(Y|X;\mathbf{w},\sigma)$ equivalent to maximizing log-likelihood

$$\begin{aligned}\log p(Y|X;\mathbf{w},\sigma) &= \log \prod_{i=1}^{N} p(y_i|\mathbf{x}_i,\mathbf{w},\sigma) \\ &= \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i,\mathbf{w},\sigma)\end{aligned}$$

Log-likelihood, Gaussian noise

$$p(y|\mathbf{x};\mathbf{w},\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y-f(\mathbf{x};\mathbf{w}))^2}{2\sigma^2}\right)$$

$$\begin{aligned}\log p(Y|X;\mathbf{w},\sigma) &= \sum_{i=1}^{N}\log p(y_i|\mathbf{x}_i,\mathbf{w},\sigma)\\ &= \sum_{i=1}^{N}\left[-\frac{(y_i-f(\mathbf{x}_i;\mathbf{w}))^2}{2\sigma^2} - \log\sigma\sqrt{2\pi}\right]\\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-f(\mathbf{x}_i;\mathbf{w}))^2 - N\log\sigma\sqrt{2\pi}.\end{aligned}$$

- Red terms are independent of $\mathbf{w}$

Maximum likelihood

- A new loss function: *log-loss* – negative conditional log-probability of the training data

$$L(f(\mathbf{x};\mathbf{w}),y) = -\log p(y|\mathbf{x};\mathbf{w},\sigma)$$

- Maximizing log-likelihood is *always* equivalent to minimizing log-loss
- Maximizing log-likelihood under the Gaussian noise model

$$\begin{aligned}\operatorname*{argmax}_{\mathbf{w}}\sum_{i=1}^{N}\log p(y_i|\mathbf{x}_i;\mathbf{w},\sigma) &= \operatorname*{argmax}_{\mathbf{w}} -\sum_{i=1}^{N}(y_i-f(\mathbf{x}_i;\mathbf{w}))^2\\ &= \operatorname*{argmin}_{\mathbf{w}}\sum_{i=1}^{N}(y_i-f(\mathbf{x}_i;\mathbf{w}))^2\end{aligned}$$

is equivalent to minimizing squared loss