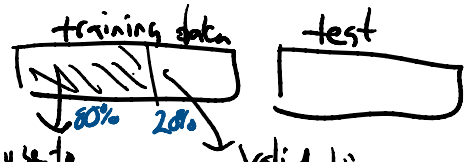
5.3-5.5: Hyperparameters and Validation Sets; Maximum Likelihood

Saturday, June 9, 2018 10:12 PM

hyperparameters that control model capacity cannot be learned in training
→ would always learn maximum possible capacity
↳ overfitting

validation set - data withheld from training
cannot use test set to decide hyperparameters



use to learn parameters

validation
guide selection of hyperparameters
estimate generalization error

benchmarks became stale
researchers build on results, so corrupted by use of valid, test data

(5.3.1) Cross-Validation

k-fold cross validation
for trial i, learn on data without subset i, test on subset i

(5.4) Estimators, Bias, Variance

(5.4.1) Point Estimation

point estimator (or **statistic**) is any function of the data:

$$\hat{\theta}_m = g\left(x^{(1)}, \ldots, x^{(m)}\right)$$

$\{x^{(1)}, \ldots, x^{(m)}\}$ set of i.i.d. points

↖ goal is to approximate θ, true parameter value

function estimation

$$y = f(x) + \epsilon$$

↑ goal is to find $\hat{f}$ that approximates f.

↖ not predictable from x

note: this is a type of point estimator, in function space

(5.4.2) Bias

... a type of point estimator, in function space

(5.4.2) Bias

$$\text{bias}(\hat{\theta}_m) = \mathbb{E}(\hat{\theta}_m) - \theta$$

expectation over the data; true value

estimator is **unbiased** if $\text{bias}(\hat{\theta}_m) = 0$

asymptotically unbiased if $\lim_{m \to \infty} \text{bias}(\hat{\theta}_m) = 0$.

$\Rightarrow \lim_{m \to \infty} \mathbb{E}(\hat{\theta}_m) = \theta$.

(potential exercises, examples in book)

(5.4.3) Variance and Standard Error

$\text{Var}(\hat{\theta})$ standard error is $\sqrt{\text{Var}(\hat{\theta})}$ or $\text{SE}(\hat{\theta})$

how much we expect the estimate to vary as we resample from underlying data-generating process

std err of mean:

$$\text{SE}(\hat{\mu}_m) = \sqrt{\text{Var}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right]} = \frac{\sigma}{\sqrt{m}}$$ where σ^2 is true variance of $x^{(i)}$

does **not** provide an unbiased estimate of standard deviation ?!

tends to underestimate it, but used in practice

often, in ML estimate generalization error as sample mean of error on the test set

central limit theorem: mean will be approx. distributed with normal distribution

e.g., 95% confidence interval around mean $\hat{\mu}_m$ is:

$$(\hat{\mu}_m - 1.96\,\text{SE}(\hat{\mu}_m),\ \hat{\mu}_m + 1.96\,\text{SE}(\hat{\mu}_m))$$

(5.4.4) Trade off Bias and Variance to Minimize MSE.

$$\text{MSE} = \mathbb{E}\left[(\hat{\theta}_m - \theta)^2\right] = \text{Bias}(\hat{\theta}_m)^2 + \text{Var}(\hat{\theta}_m).$$

overall expected deviation; estimator; true

overall expected deviation — estimator true

increasing capacity tends to: decrease bias, increase variance

(5.4.5) Consistency

$\text{plim}_{m \to \infty} \hat{\theta}_m = \theta$ — consistency ("weak")

convergence in probability

$P(|\hat{\theta}_m - \theta| > \epsilon) \to 0$ as $m \to \infty$

consistency ensures bias induced by estimator diminishes as m grows

(5.5) Maximum Likelihood Estimation

how can we derive good estimators?

functions that are

m examples drawn i.i.d. from $p_{data}(x)$

$\mathbb{X} = \{x^{(1)}, \ldots, x^{(m)}\}$

$p_{model}(x; \theta)$ - parametric family of probability distributions over space indexed by θ

so, $p_{model}(x; \theta)$ maps $x \to$ real number estimating true probability $p_{data}(x)$.

maximum likelihood estimator for θ is defined as:

$$\theta_{ML} = \arg\max_\theta p_{model}(\mathbb{X}; \theta) = \arg\max_\theta \prod_{i=1}^{m} p_{model}(x^{(i)}; \theta)$$

difficult to compute accurately (underflow)

so, take log: (doesn't change arg max):

$$\theta_{ML} = \arg\max_\theta \sum_{i=1}^{m} \log p_{model}(x^{(i)}; \theta).$$

also, doesn't change when scaled $\Rightarrow$ expectation

$$\theta_{ML} = \arg\max_\theta \mathbb{E}_{x \sim \hat{p}_{data}} \log p_{model}(x; \theta). \quad (5.59)$$

intuition: minimize dissimilarity between $\hat{p}_{data}$ and p_{model}.

intuition: minimize dissimilarity between $\hat{p}_{data}$ and p_{model}.
empirical dist. defined by training set; model distribution

KL divergence:

$$D_{KL}(\hat{p}_{data} \| p_{model}) = \mathbb{E}_{x \sim \hat{p}_{data}}\left[\log \hat{p}_{data}(x) - \log p_{model}(x)\right]$$

$\hat{p}_{data}$: data generating process

training a model to minimize KL divergence means minimize

$$-\mathbb{E}_{x \sim \hat{p}_{data}}\left[\log p_{model}(x)\right]$$

(same as maximizing 5.59)

corresponds to minimizing cross-entropy between distributions

(5.5.1) Conditional Log-Likelihood and Mean Squared Error

generalize Maximum Likelihood Estimation to estimate conditional probability, $P(y \mid x; \theta)$. (predict y given x)

X: all inputs, Y: all targets

Conditional MLE:

$$\theta_{ML} = \arg\max_{\theta} P(Y \mid X; \theta)$$

assuming i.i.d.,

$$\theta_{ML} = \arg\max_{\theta} \sum_{i=1}^{m} \log P(y^{(i)} \mid x^{(i)}; \theta)$$

Example: Linear Regression as Max Likelihood

explains why minimizing mean squared error

(5.5.2) Properties of Maximum Likelihood

main appeal: best estimator asymptotically as $m \to \infty$.
rate of convergence

satisfies consistency: (converges to true value) under 2 conditions:

- p_{data} must lie within model family $p_{model}(\cdot; \theta)$
  (otherwise, no estimator can recover p_{data})

- p_{data} ... within model family $p_{model}(\cdot\,; \theta)$
  (otherwise, no estimator can recover p_{data})
- true dist. of p_{data} must correspond to <u>exactly one</u> value of θ.
  (otherwise, can recover, but can't know which θ is used)