

CS195-5 : Introduction to Machine Learning

Lecture 8

Greg Shakhnarovich

September 22 2006

Revised October 24th, 2006

Announcements

Review

- Bayes classifier, that achieves minimal risk (Bayes risk):

$$h^*(\mathbf{x}) = \operatorname{argmax}_c p(y = c | \mathbf{x}) = \operatorname{argmax}_c p(\mathbf{x} | y = c) P_c$$

- Discriminant analysis: define $\delta_c \triangleq \log p(\mathbf{x} | c) P_c$, then

$$h^*(\mathbf{x}) = \operatorname{argmax}_c \delta_c(\mathbf{x}).$$

- 2 Gaussians, equal covariances: linear discriminant analysis (LDA).
- 2 Gaussians, general covariances: quadratic discriminant analysis.
- Generative models: assume a parametric form $p(\mathbf{x}|c; \theta)$, estimate θ from data, derive Bayes classifier pretending the estimated distribution is correct.

Generative models

- Operate under the “pretense” that $\mathbf{x} \sim p(\mathbf{x}; \theta)$.
- Under that assumption we depend on the estimation to infer the parameters θ .
- Examples we have seen so far:
 - The additive noise model for regression.
 - Modeling class-conditionals by Gaussians in classification.

Estimation theory

- An *estimator* $\hat{\theta}_N$ of a parameter θ is a function that takes the data $X_N = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ and produces an estimated value $\hat{\theta}$.
 - Estimator $\hat{\theta}_N$ is a procedure; an *estimate* $\hat{\theta}$ is a value obtained by that procedure.
 - E.g., a maximum likelihood estimator for a 1D Gaussian mean, given X_N , produces an estimate (number) $\hat{\mu}$.
- The estimate $\hat{\theta}$ is a random variable since it is based on a randomly drawn set X_N .
- We can talk about $E[\hat{\theta}|X_N]$ and $\text{var}(\hat{\theta}|X_N)$.
(When θ is a vector, we have $\text{Cov}(\hat{\theta})$.)
 - *Analysis done assuming that the data is distributed according to $p(\mathbf{x}; \theta)$!*

Bias of an estimator

- The *bias* of an estimator $\hat{\theta}_N$ is defined as

$$\text{bias}(\hat{\theta}_N) \triangleq E_{X_N} [\hat{\theta}_N - \theta].$$

i.e. the expected deviation of the estimate from the correct parameter (taken over all possible sets of N examples).

- An *unbiased* estimator therefore satisfies $E_{X_N} [\hat{\theta}_N] = \theta$.

- Example: ML estimators of 1D Gaussian parameters

$$\hat{\mu}_{ML} = \frac{1}{N} \sum_i x_i, \quad \hat{\sigma}^2_{ML} = \frac{1}{N} \sum_i (x_i - \hat{\mu})^2.$$

Bias of an estimator

- The *bias* of an estimator $\hat{\theta}_N$ is defined as

$$\text{bias}(\hat{\theta}_N) \triangleq E_{X_N} [\hat{\theta}_N - \theta].$$

i.e. the expected deviation of the estimate from the correct parameter (taken over all possible sets of N examples).

- An *unbiased* estimator therefore satisfies $E_{X_N} [\hat{\theta}_N] = \theta$.

- Example: ML estimators of 1D Gaussian parameters

$$\hat{\mu}_{ML} = \frac{1}{N} \sum_i x_i, \quad \hat{\sigma}^2_{ML} = \frac{1}{N} \sum_i (x_i - \hat{\mu})^2.$$

- Turns out $\hat{\mu}$ is unbiased;

Bias of an estimator

- The *bias* of an estimator $\hat{\theta}_N$ is defined as

$$\text{bias}(\hat{\theta}_N) \triangleq E_{X_N} [\hat{\theta}_N - \theta].$$

i.e. the expected deviation of the estimate from the correct parameter (taken over all possible sets of N examples).

- An *unbiased* estimator therefore satisfies $E_{X_N} [\hat{\theta}_N] = \theta$.

- Example: ML estimators of 1D Gaussian parameters

$$\hat{\mu}_{ML} = \frac{1}{N} \sum_i x_i, \quad \hat{\sigma}^2_{ML} = \frac{1}{N} \sum_i (x_i - \hat{\mu})^2.$$

- Turns out $\hat{\mu}$ is unbiased; however, $\hat{\sigma}_{ML}$ *underestimates* the variance in the data!

$$E [\hat{\sigma}^2_{ML}] = \frac{N-1}{N} \sigma^2.$$

Consistency of an estimator

- If we have enough data, bias may not be so much of a problem.
- An estimator $\hat{\theta}_N$ is *consistent* if

$$\lim_{N \rightarrow \infty} \hat{\theta}_N = \theta.$$

Note: this limit is *in probability*.

- So, $\hat{\sigma}_{ML}^2 = \frac{1}{N} \sum_i (x_i - \mu_{ML})^2$, even though biased, is a consistent estimator of σ^2 .

Bias-variance dilemma

- We can associate squared loss with the error $\hat{\theta} - \theta$.
- Denote $\bar{\theta}_N = E[\hat{\theta}_N]$. Then, the expected error:

$$\begin{aligned} E[(\hat{\theta}_N - \theta)^2] &= E[(\hat{\theta}_N - \bar{\theta}_N + \bar{\theta}_N - \theta)^2] \\ &= E[(\hat{\theta}_N - \bar{\theta}_N)^2] + 2(\bar{\theta}_N - \theta)E[\hat{\theta}_N - \bar{\theta}_N] + E[(\bar{\theta}_N - \theta)^2] \\ &= (\bar{\theta}_N - \theta)^2 + E[(\hat{\theta}_N - \bar{\theta}_N)^2] \\ &= \text{bias}^2(\hat{\theta}_N) + \text{var}(\hat{\theta}_N). \end{aligned}$$

- This is the same phenomenon that we saw for regression:
 - The bias^2 term corresponds to structural error of the model,
 - the variance is the approximation error due to finite data.

Data complexity of estimation

- Two kinds of complexity are relevant to learning/estimation algorithms:
 - Computational (time) complexity: more data \Rightarrow higher cost

Data complexity of estimation

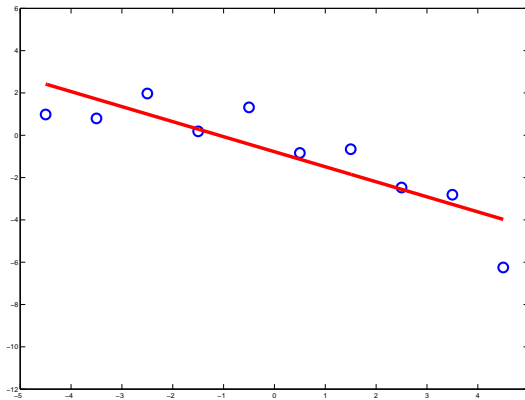
- Two kinds of complexity are relevant to learning/estimation algorithms:
 - Computational (time) complexity: more data \Rightarrow higher cost
 - Data complexity: *less* data \Rightarrow higher cost.
- Consider the ML estimate of the Gaussian in \mathbb{R}^d .
 - mean: need to fit d parameters.
 - Covariance:

Data complexity of estimation

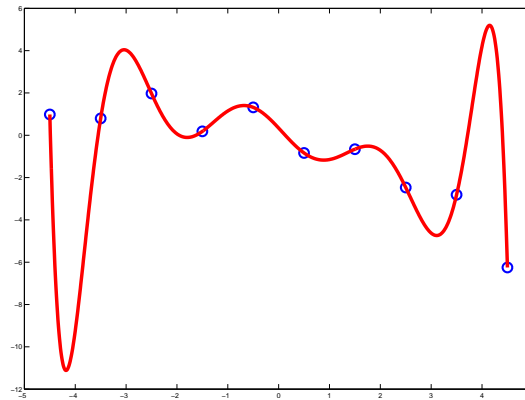
- Two kinds of complexity are relevant to learning/estimation algorithms:
 - Computational (time) complexity: more data \Rightarrow higher cost
 - Data complexity: *less* data \Rightarrow higher cost.
- Consider the ML estimate of the Gaussian in \mathbb{R}^d .
 - mean: need to fit d parameters.
 - Covariance: $d + d(d - 1)/2$ parameters.
- Rule of thumb: need 10-30 examples *per parameter*.

Model complexity

- Intuitively, the complexity of the model can be measured by the number of “degrees of freedom” (independent parameters).
 - The more complex the model, the more data we need to fit it
⇒ For a given number of points, a more complex model is more likely to overfit.
 - Example: polynomial regression of order m .



$m = 1$, 2 parameters



$m = 10$, 11 parameters

- This is an issue only because of finite training data!

Dealing with model complexity

- We have already seen one way to deal with overfitting: cross-validation.
- Another way is to restrict the complexity of the model.
- For the case of d -variate Gaussian, we can restrict the covariance matrix. We need to estimate d parameters for μ plus:

$$\text{Model } \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{1d} & \cdots & \cdots & \sigma_d^2 \end{bmatrix}$$

# param.	$d + d(d - 1)/2$
----------	------------------

Dealing with model complexity

- We have already seen one way to deal with overfitting: cross-validation.
- Another way is to restrict the complexity of the model.
- For the case of d -variate Gaussian, we can restrict the covariance matrix. We need to estimate d parameters for μ plus:

$$\text{Model } \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2d} \\ \dots & \dots & \dots & \dots \\ \sigma_{1d} & \dots & \dots & \sigma_d^2 \end{bmatrix} \quad \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \sigma_d^2 \end{bmatrix}$$

# param.	$d + d(d - 1)/2$	d
----------	------------------	-----

Dealing with model complexity

- We have already seen one way to deal with overfitting: cross-validation.
- Another way is to restrict the complexity of the model.
- For the case of d -variate Gaussian, we can restrict the covariance matrix. We need to estimate d parameters for μ plus:

$$\text{Model } \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1d} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2d} \\ \dots & \dots & \dots & \dots \\ \sigma_{1d} & \dots & \dots & \sigma_d^2 \end{bmatrix} \quad \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \sigma_d^2 \end{bmatrix} \quad \sigma^2 \mathbf{I}$$

# param.	$d + d(d - 1)/2$	d	1
----------	------------------	-----	-----