CS 446/ECE 449: Machine Learning

Lecture 9: PAC Learning Theory (I)

Han Zhao
02/13/2024

ILLINOIS
Computer Science
GRAINGER COLLEGE OF ENGINEERING

Recap: Supervised Learning Algorithms

Models we have learned so far:

Model	Linear?	Parametric?	Loss	Generative/ Discriminative
K-nearest neighbor	N	N	N/A	Discriminative
Naive Bayes	Y	Y	NLL	Generative
Logistic regression	Y	Y	Logistic/NLL	Discriminative
Linear SVM	Y	Y	Hinge	Discriminative
Kernelized SVM	N	N	Hinge	Discriminative
Decision Tree	N	N	N/A	Discriminative
AdaBoost	N	N	Exp	Discriminative

Note:

- NLL = negative log-likelihood
- Generative = modeling $\Pr(X, Y)$
- Discriminative = modeling $\Pr(Y \mid X)$

Lecture Today

- Bayes Error, Bayes Predictor
- Error Decomposition

Bayes Error Rate

So far we have learned many different classification algorithms. Beyond their different design choices, how should we compare their performance theoretically?

- For a given prediction problem, what is the optimal error that we can hope to achieve? Which predictor will achieve the optimal error?
- Given a problem and a model, how far is our model from the optimal predictor?

Bayes Error Rate

The learning process:

- We can choose a predictor f from some pre-defined class of functions $\mathscr{F}$, e.g., the class of linear predictors, decision trees, kernel machines, neural networks, etc.

We also have our training data $\mathscr{D} := \{(x^{(i)}, y^{(i)})\}_{i=1}^{n} \sim \mu$ sampled independently and identically (iid) from the underlying distribution μ over $\mathscr{X} \times \mathscr{Y}$

We can then talk about two error measures (classification):

$$\text{Training error: } \hat{\varepsilon}_{\mathscr{D}}(f) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(f(x^{(i)}) \neq y^{(i)})$$

$$\text{Test error: } \varepsilon_{\mu}(f) := \mathbb{E}_{\mu}\left[\mathbb{I}(f(X) \neq Y)\right] = \Pr_{\mu}(f(X) \neq Y)$$

We are interested in finding f that minimizes the test error but we can only observe the training error

Bayes Error Rate

Bayes error rate: the theoretically minimum test error that can be achieved:

$$\text{Bayes error: } \varepsilon_\mu^* := \inf_{f:\mathcal{X}\to\mathcal{Y}} \varepsilon_\mu(f)$$

Assuming X is a continuous RV and let $p(x)$ be the probability density of X. Then for any classifier $f: \mathcal{X} \to \{0,1\}$, we have:

$$\begin{aligned}
\varepsilon_\mu(f) &= \Pr_\mu(f(X) \neq Y) \\
&= \int_{\mathcal{X}} \left(\Pr_\mu(Y=1 \mid X=x) \cdot \mathbb{I}(f(x)=0) + \Pr_\mu(Y=0 \mid X=x) \cdot \mathbb{I}(f(x)=1) \right) p(x)\, \mathrm{d}x \\
&\geq \int_{\mathcal{X}} \min \left\{ \Pr_\mu(Y=1 \mid X=x), \Pr_\mu(Y=0 \mid X=x) \right\} p(x)\, \mathrm{d}x \\
&= \mathbb{E}_\mu \left[\min \left\{ \Pr_\mu(Y=1 \mid X), \Pr_\mu(Y=0 \mid X) \right\} \right] \\
&= \frac{1}{2} - \frac{1}{2} \mathbb{E}_\mu \left[\,|\, 2\eta(X) - 1 \,|\, \right]
\end{aligned}$$

where $\eta(X) := \Pr_\mu(Y=1 \mid X)$ is the conditional probability

Bayes Error Rate

Bayes error depends on the distribution μ:

$$
\begin{aligned}
\varepsilon_\mu(f) &= \Pr_\mu(f(X) \neq Y) \\
&= \int_{\mathcal{X}} \left(\Pr_\mu(Y = 1 \,|\, X = x) \cdot \mathbb{I}(f(x) = 0) + \Pr_\mu(Y = 0 \,|\, X = x) \cdot \mathbb{I}(f(x) = 1) \right) p(x)\ \mathrm{d}x \\
&\geq \int_{\mathcal{X}} \min \left\{ \Pr_\mu(Y = 1 \,|\, X = x), \Pr_\mu(Y = 0 \,|\, X = x) \right\} p(x)\ \mathrm{d}x \\
&= \mathbb{E}_\mu \left[\min \left\{ \Pr_\mu(Y = 1 \,|\, X), \Pr_\mu(Y = 0 \,|\, X) \right\} \right] \\
&= \boxed{\frac{1}{2} - \frac{1}{2} \mathbb{E}_\mu \left[\,|\, 2\eta(X) - 1 \,|\, \right]}
\end{aligned}
$$

Bayes error rate: ε_μ^*

- The Bayes error only depends on the distribution μ
- It's unknown since we don't know μ in practice
- It's always ≤ 0.5

Bayes Error Rate

The classifier that achieves the Bayes error is called the Bayes classifier, and it has the following form:

$$\boxed{\eta(X) := \Pr(Y = 1 \mid X)}$$

$$f_{\text{Bayes}}(X) := \begin{cases} 1 & \text{if } \eta(X) \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

- Again, this is unknown since we don't know μ
- Recall the proof:

$$\begin{aligned} \varepsilon_\mu(f) &= \int_{\mathcal{X}} \left(\Pr_\mu(Y = 1 \mid X = x) \cdot \mathbb{I}(f(x) = 0) + \Pr_\mu(Y = 0 \mid X = x) \cdot \mathbb{I}(f(x) = 1) \right) p(x)\ \mathrm{d}x \\ &\geq \int_{\mathcal{X}} \min \left\{ \Pr_\mu(Y = 1 \mid X = x), \Pr_\mu(Y = 0 \mid X = x) \right\} p(x)\ \mathrm{d}x \\ &= \frac{1}{2} - \frac{1}{2}\mathbb{E}_\mu \left[\left| 2\eta(X) - 1 \right| \right] \end{aligned}$$

Think: when will $\varepsilon_\mu^* = 0$? when will $\varepsilon_\mu^* = 0.5$?

Bayes Error Rate

Intuitively, the Bayes error is a measure of the “noise” in the underlying distribution:

Bayes error rate: $\varepsilon_\mu^* = \frac{1}{2} - \frac{1}{2}\mathbb{E}\left[\,|\,2\eta(X) - 1\,|\,\right]$

$$\boxed{\eta(X) := \Pr(Y = 1 \mid X)}$$

- If $\forall x, \eta(x) = 1$, or $\eta(x) = 0$, then $\varepsilon_\mu^* = 0$
- If $\forall x, \eta(x) = \frac{1}{2}$, then $\varepsilon_\mu^* = \frac{1}{2}$

Bayes Error Rate

Intuitively, the Bayes error is a measure of the "noise" in the underlying distribution:

Bayes error rate: $\varepsilon_\mu^* = \frac{1}{2} - \frac{1}{2}\mathbb{E}\left[|2\eta(X) - 1|\right]$

$$\boxed{\eta(X) := \Pr(Y = 1 \mid X)}$$

Example: Suppose we have the following data generative process. There exists a vector w^*, such that for each x, the labels are generated in the following process:

- First, compute the label $y = \text{sgn}(w^{*\top} x)$
- Then, with probability $0 < p < 0.5$, flip the label y

Question: What's the Bayes error rate for this example?

Bayes Error Rate

The concept is not unique to classification problems. For regression problems, under the squared loss:

$$
\begin{aligned}
\forall f,\ \varepsilon_\mu(f) &= \mathbb{E}_\mu\left[(f(X)-Y)^2\right] \\
&= \mathbb{E}_X\mathbb{E}_Y\left[(f(X)-Y)^2 \,|\, X\right] \\
&= \mathbb{E}_X\mathbb{E}_Y\left[(f(X)-\mathbb{E}[Y|X]+\mathbb{E}[Y|X]-Y)^2 \,|\, X\right] \\
&= \mathbb{E}_X\mathbb{E}_Y\left[(f(X)-\mathbb{E}[Y|X])^2+(\mathbb{E}[Y|X]-Y)^2\right. \\
&\quad \left. + \cancel{2(f(X)-\mathbb{E}[Y|X])(\mathbb{E}[Y|X]-Y)} \,|\, X\right] \\
&\geq \mathbb{E}_X\mathbb{E}_Y\left[(\mathbb{E}[Y|X]-Y)^2 \,|\, X\right]
\end{aligned}
$$

Bayes error rate: $= \mathbb{E}_X \mathrm{Var}[Y|X]$

Bayes optimal regressor: $f_{\text{Bayes}}(X) = \mathbb{E}[Y|X]$

Bayes Error Rate

Again, the Bayes error in regression can also be understood as a measure of the “noise” in the underlying distribution:

Bayes error rate: $\varepsilon_\mu^* = \mathbb{E}\mathrm{Var}[Y|X]$

Example: Suppose we have the following data generative process. There exists a vector w^*, such that for each x, the labels are generated in the following process:

- First, compute the label $y = w^{*\top} x$
- Then, inject a white noise $\epsilon \sim \mathcal{N}(0, \delta^2)$ into the label so that $y \leftarrow y + \epsilon$

Question: What’s the Bayes error under the squared loss for this example?

Bayes Error Rate

The optimal error a learner can hope to achieve also depends on the class of functions $\mathscr{F}$ it can choose from, called hypothesis class

For binary classification problems:

- If $\mathscr{F}$ contains all the binary functions, then $\inf_{f \in \mathscr{F}} \varepsilon_\mu(f) = \varepsilon^*_\mu$
- If $\mathscr{F}$ is very restricted, e.g., only contains constant functions, then $\inf_{f \in \mathscr{F}} \varepsilon_\mu(f) = \min\{\Pr(Y=0), \Pr(Y=1)\}$

Clearly, in the second case, the error is larger than the Bayes error.

Bayes Error Rate

The optimal error a learner can hope to achieve also depends on the class of functions $\mathscr{F}$ it can choose from, called hypothesis class

For regression problems under mean-squared error:

- If $\mathscr{F}$ contains all the real-valued functions, then $\inf_{f \in \mathscr{F}} \varepsilon_\mu(f) = \varepsilon_\mu^*$
- If $\mathscr{F}$ is very restricted, e.g., only contains constant functions, then $\inf_{f \in \mathscr{F}} \varepsilon_\mu(f) = \mathrm{Var}[Y]$

In the second case, the error is larger than the Bayes error by the law of total variance: $\mathrm{Var}[Y] = \mathbb{E}\mathrm{Var}[Y \mid X] + \mathrm{Var}\mathbb{E}[Y \mid X] \geq \mathbb{E}\mathrm{Var}[Y \mid X]$

Bayes Error Rate

Summary:

Binary classification:

Bayes error rate: $\varepsilon_\mu^* = \mathbb{E}\min\left\{\Pr(Y=1\,|\,X), \Pr(Y=0\,|\,X)\right\}$

Bayes optimal classifier: $f_{\text{Bayes}}(X) := \begin{cases} 1 & \text{if } \Pr(Y=1|X) \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$

Regression with squared loss:

Bayes error rate: $\varepsilon_\mu^* = \mathbb{E}\mathrm{Var}[Y\,|\,X]$

Bayes optimal regressor: $f_{\text{Bayes}}(X) = \mathbb{E}[Y\,|\,X]$

Error Decomposition

For a given hypothesis class $\mathscr{F}$, we may have $f_{\text{Bayes}} \notin \mathscr{F}$. In this case we cannot hope to achieve ε_μ^*, but instead $\inf\limits_{f \in \mathscr{F}} \varepsilon_\mu(f)$.

Error decomposition: $\forall f \in \mathscr{F}$:

$$\varepsilon_\mu(f) = \boxed{\varepsilon_\mu(f) - \inf_{f \in \mathscr{F}} \varepsilon_\mu(f)} + \boxed{\inf_{f \in \mathscr{F}} \varepsilon_\mu(f) - \varepsilon_\mu^*} + \varepsilon_\mu^*$$

Bayes error
(depending on the inherent noise in the data)

Estimation error
(depending on the size of our data and $\mathcal{F}$)

Approximation error
(depending on the expressiveness of $\mathcal{F}$)

Error Decomposition

Error decomposition: $\forall f \in \mathscr{F}$:

$$\varepsilon_\mu(f) = \boxed{\varepsilon_\mu(f) - \inf_{f \in \mathscr{F}} \varepsilon_\mu(f)} + \boxed{\inf_{f \in \mathscr{F}} \varepsilon_\mu(f) - \varepsilon_\mu^*} + \varepsilon_\mu^*$$

Estimation error
(depending on the size of our data and $\mathcal{F}$)

Approximation error
(depending on the expressiveness of $\mathcal{F}$)

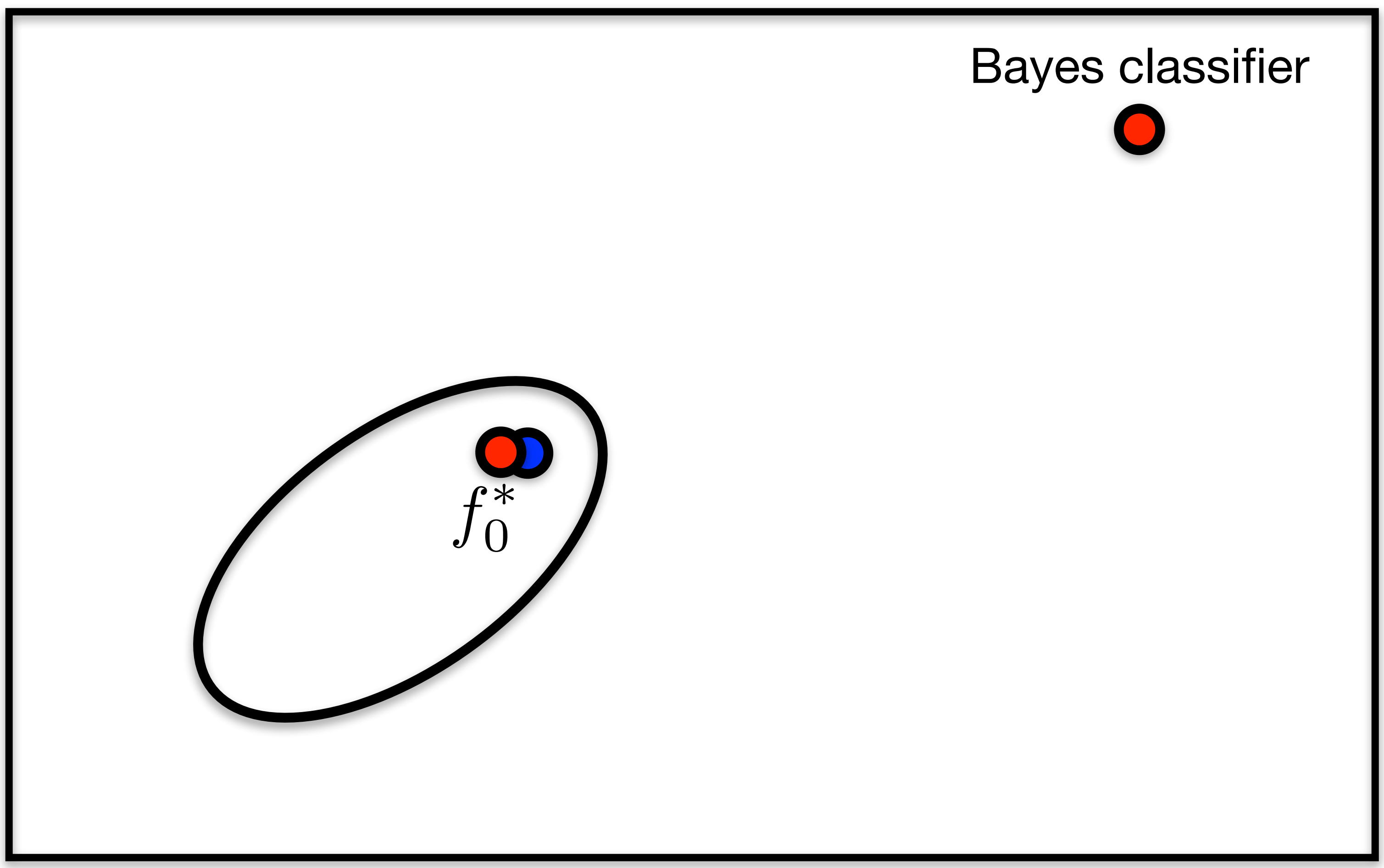
Bayes error
(depending on the inherent noise in the data)

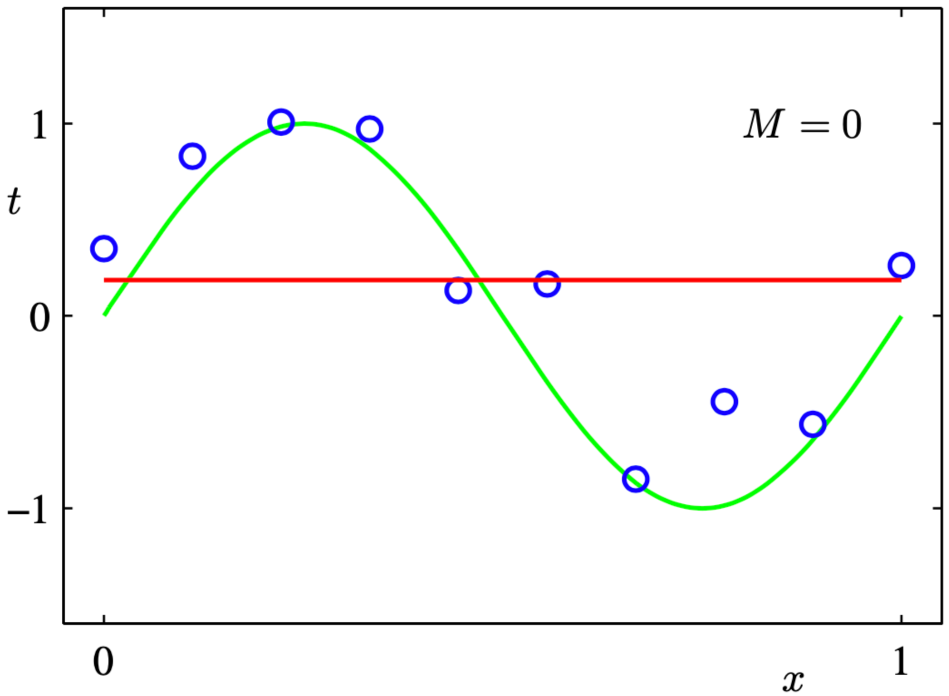
- Often the case, there is a trade-off between the estimation error and the approximation error
- If $\mathscr{F}$ is more expressive, then the approximation error gets smaller but the estimation error gets larger
- If $\mathscr{F}$ is more restricted, then the approximation error gets larger but the estimation error gets smaller (assume the size of training data is fixed)

Error Decomposition

Example: fitting a trigonometric function with polynomials (degree = M)

Degree 0

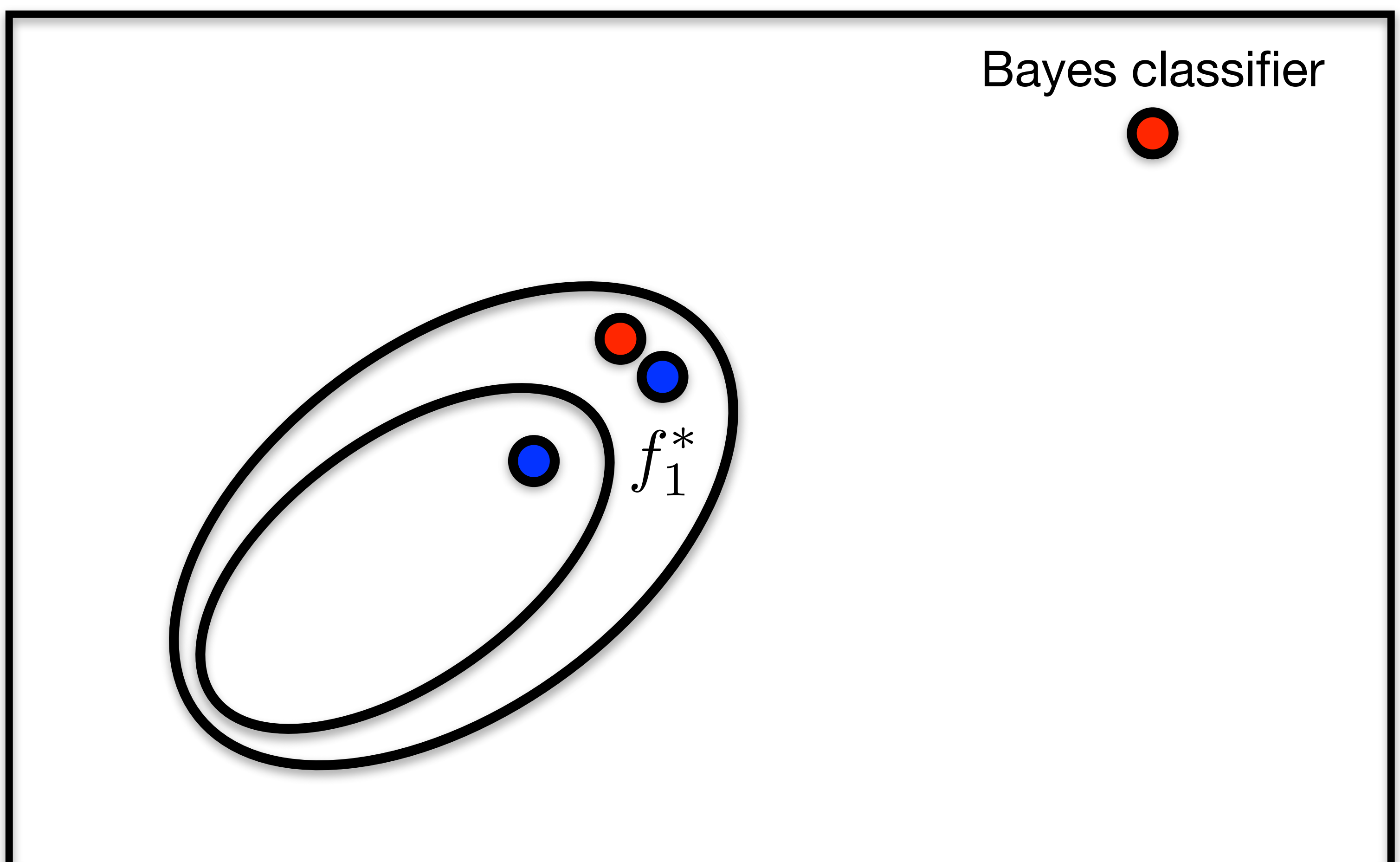


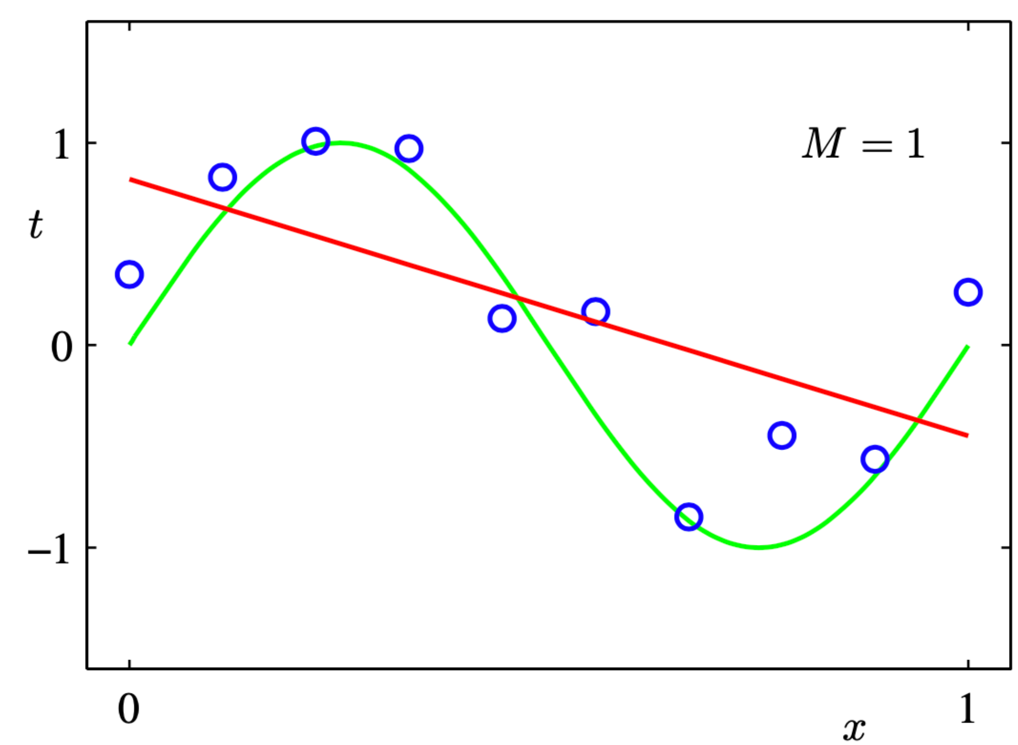


Error Decomposition

Example: fitting a trigonometric function with polynomials (degree = M)

Degree 1

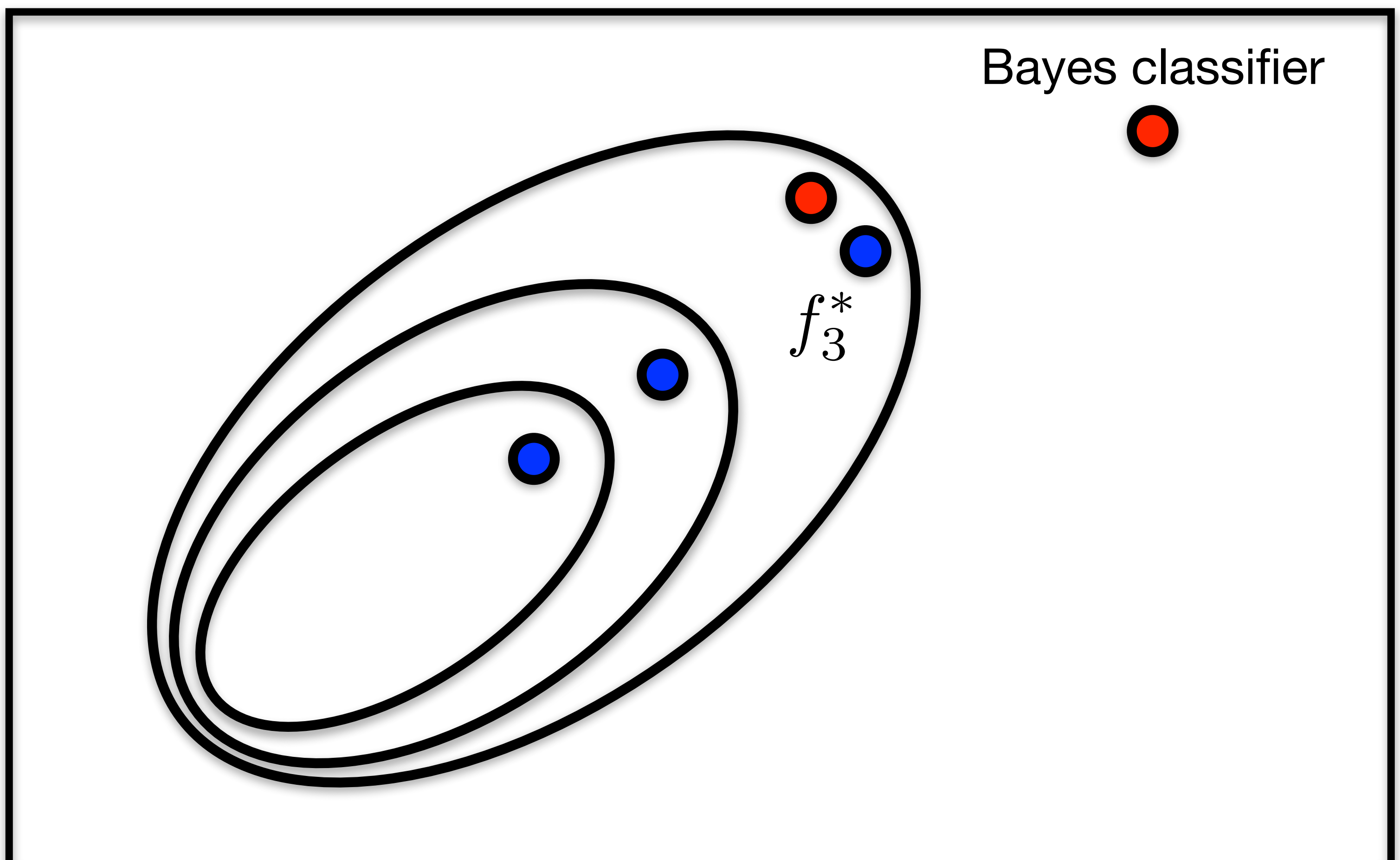


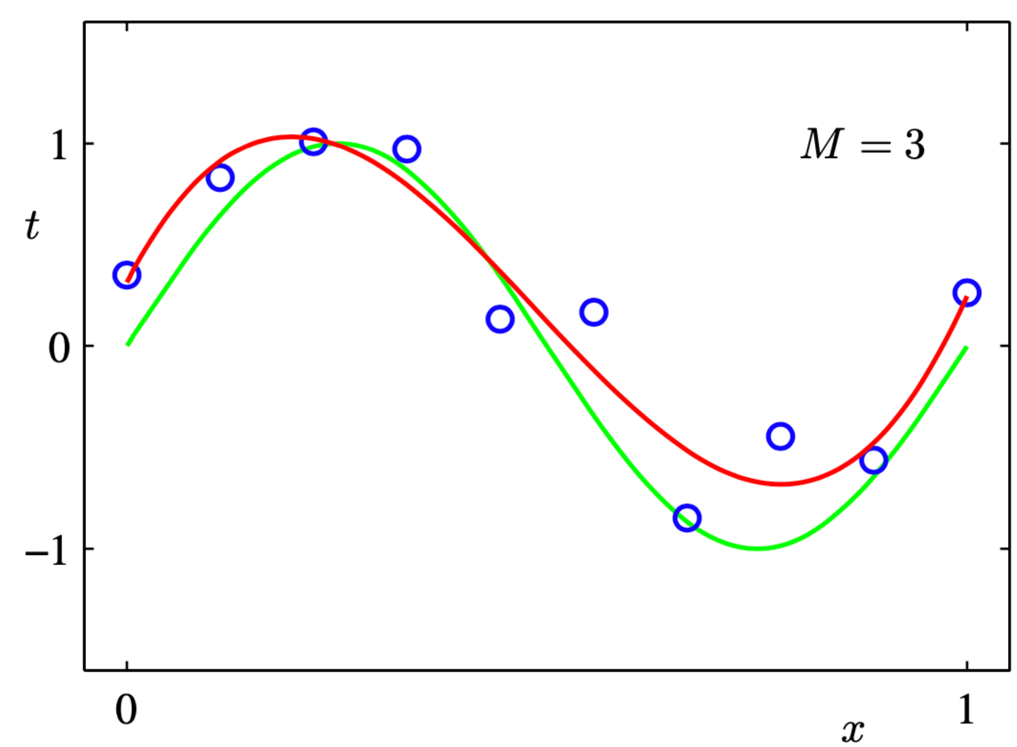


Error Decomposition

Example: fitting a trigonometric function with polynomials (degree = M)

Degree 3

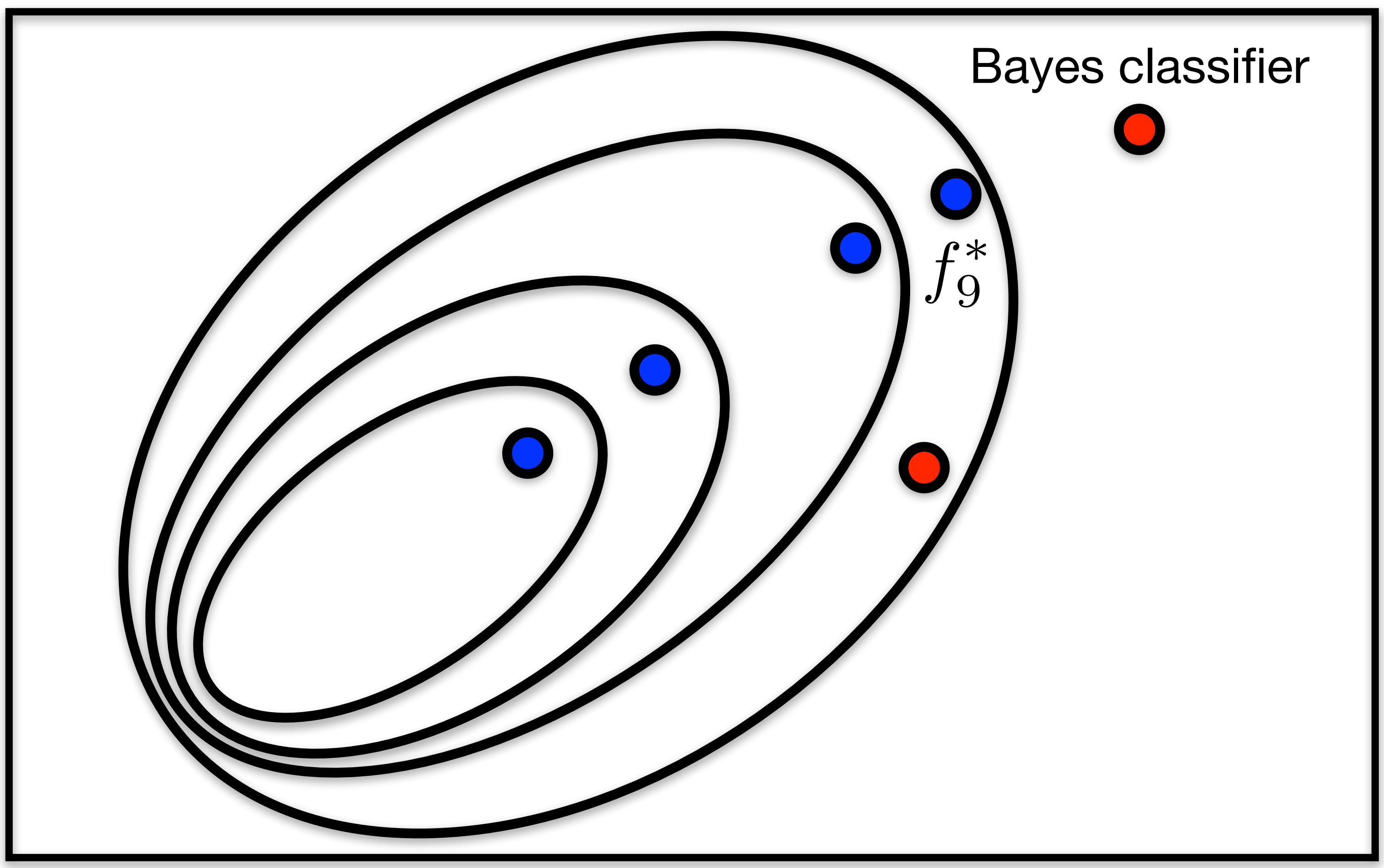


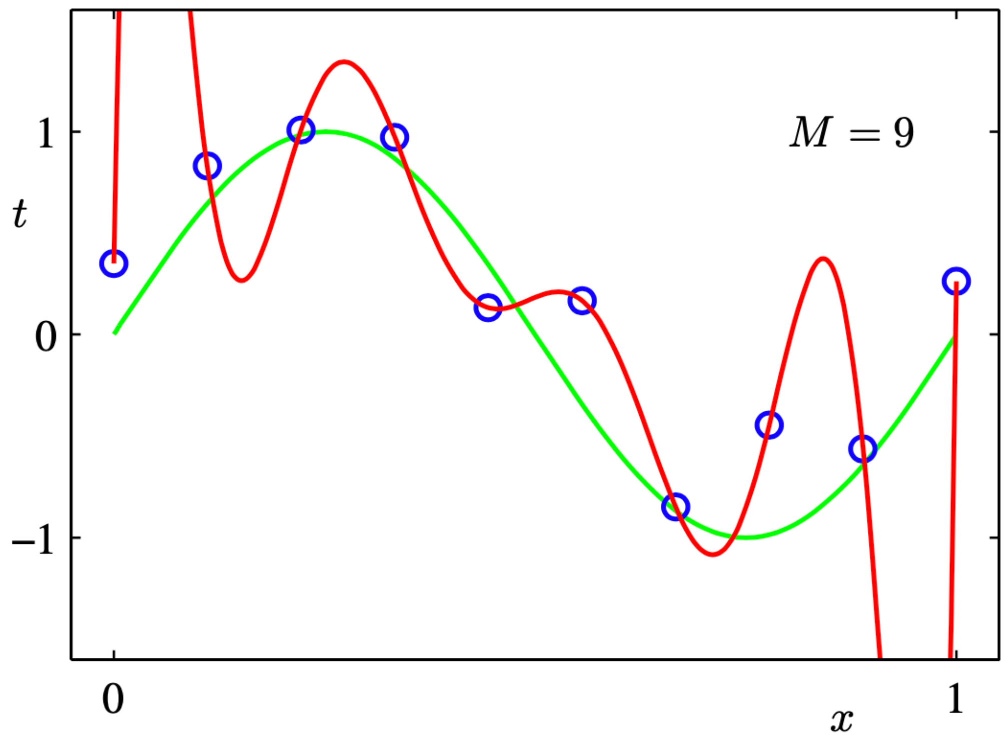


Error Decomposition

Example: fitting a trigonometric function with polynomials (degree = M)

Degree 9

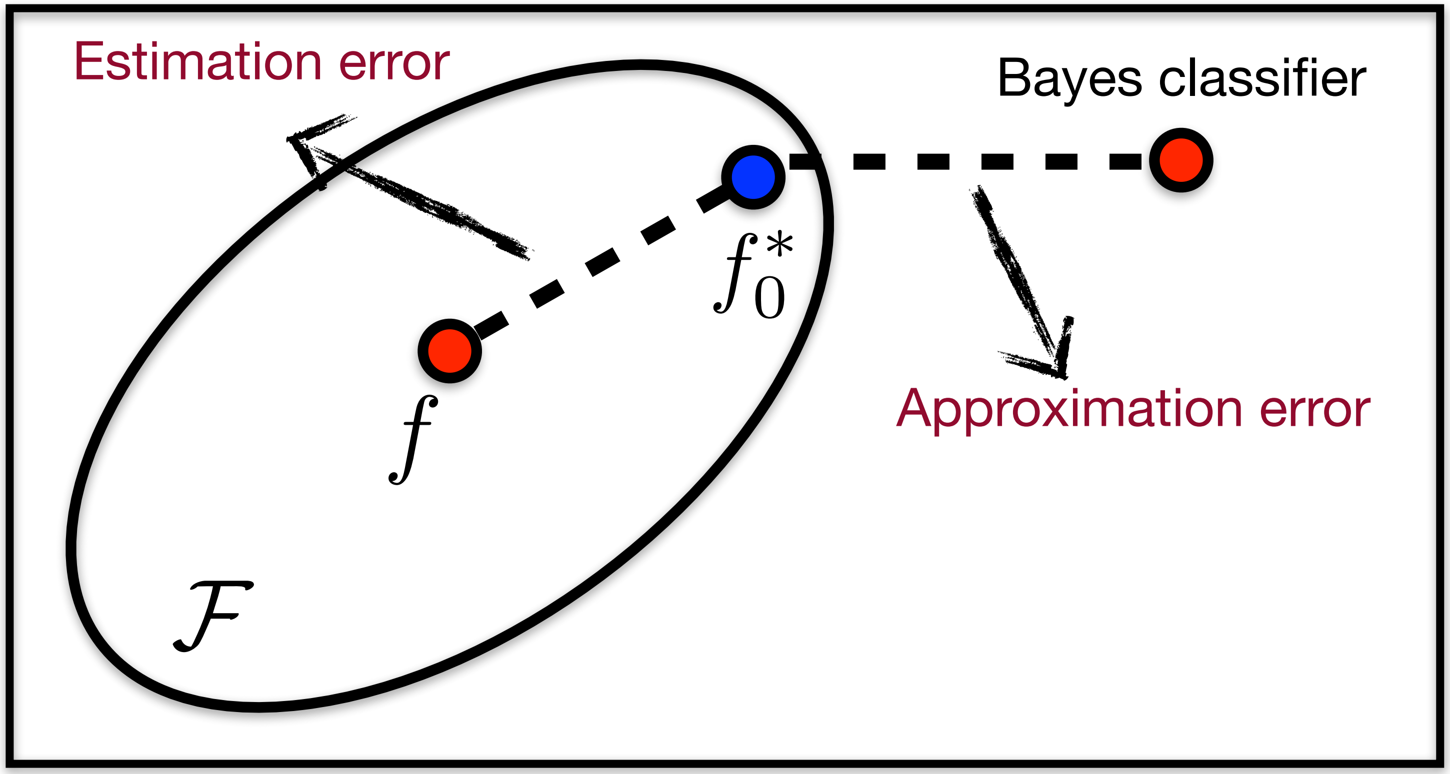




Error Decomposition

Reminder: the approximation error only depends on $\mathscr{F}$ while the estimation error depends on both $\mathscr{F}$ and data

- We should aim to minimize the estimation error
- How does the estimation error depend on the sample size, the expressiveness/richness of $\mathscr{F}$, or the distribution μ?
- Ideally, for a fixed hypothesis class $\mathscr{F}$, could we ensure that the estimation error goes to 0 as the sample size n increases?



Next Time

- Probably Approximately Correct (PAC) framework
- High-probability generalization bound
- Vapnik–Chervonenkis dimension (VC dim)