Review: Logistic regression, Gaussian naïve Bayes, linear regression, and their connections

New: Bias-variance decomposition, bias-variance tradeoff, overfitting, regularization, and feature selection

Yi Zhang
10-701, Machine Learning, Spring 2011
February 3rd, 2011

Parts of the slides are from previous 10-701 lectures

Outline

- Logistic regression
- Decision surface (boundary) of classifiers
- Generative vs. discriminative classifiers
- Linear regression
- Bias-variance decomposition and tradeoff
- Overfitting and regularization
- Feature selection

Outline

- Logistic regression
  - Model assumptions: P(Y|X)
  - Decision making
  - Estimating the model parameters
  - Multiclass logistic regression
- Decision surface (boundary) of classifiers
- Generative vs. discriminative classifiers
- Linear regression
- Bias-variance decomposition and tradeoff
- Overfitting and regularization
- Feature selection

Logistic regression: assumptions

- Binary classification
  - **f**: $X = (X_1, X_2, \ldots X_n) \rightarrow Y \in \{0, 1\}$

- Logistic regression: assumptions on P(Y|X):

$$P(Y=0|\mathbf{X},\mathbf{w}) = \frac{1}{1+exp(w_0+\sum_i w_i X_i)}$$

  - And thus:

$$P(Y=1|\mathbf{X},\mathbf{w}) = \frac{exp(w_0+\sum_i w_i X_i)}{1+exp(w_0+\sum_i w_i X_i)}$$

$$\left(= \frac{1}{1+exp(-w_0-\sum_i w_i X_i)}\right)$$

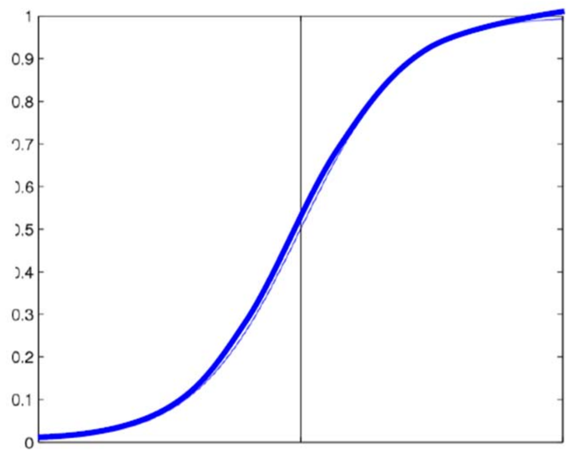
Logistic regression: assumptions

- Model assumptions: the form of P(Y|X)

$$P(Y=0|\mathbf{X},\mathbf{w}) = \frac{1}{1+exp(w_0+\sum_i w_i X_i)}$$

- "Logistic" regression
  - P(Y|X) is the ***logistic function*** applied to a linear function of X

$$\frac{1}{1+exp(-z)}$$



Decision making

- Given a logistic regression **w** and an **X**:

$$P(Y=0|\mathbf{X},\mathbf{w}) = \frac{1}{1+exp(w_0+\sum_i w_i X_i)}$$

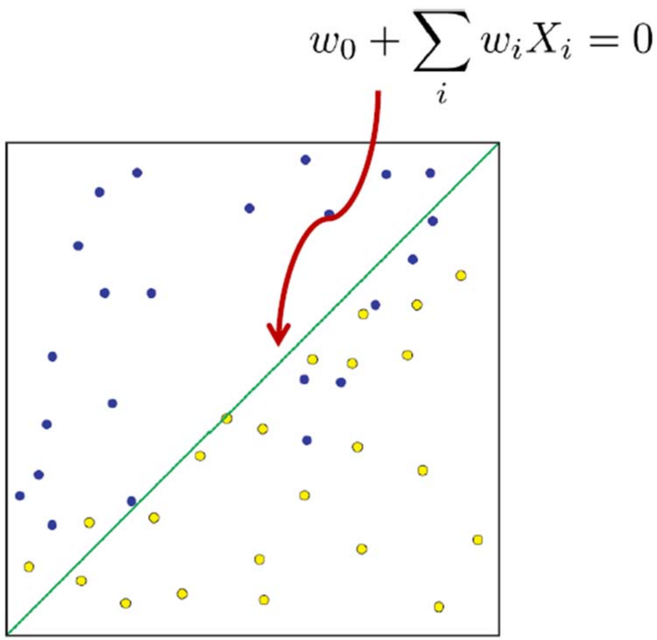
$$P(Y=1|\mathbf{X},\mathbf{w}) = \frac{exp(w_0+\sum_i w_i X_i)}{1+exp(w_0+\sum_i w_i X_i)}$$

- Decision making on Y:

$$P(Y=0|X) \underset{1}{\overset{0}{\gtrless}} P(Y=1|X)$$

Linear decision boundary !

$$0 \underset{1}{\overset{0}{\gtrless}} w_0 + \sum_i w_i X_i$$

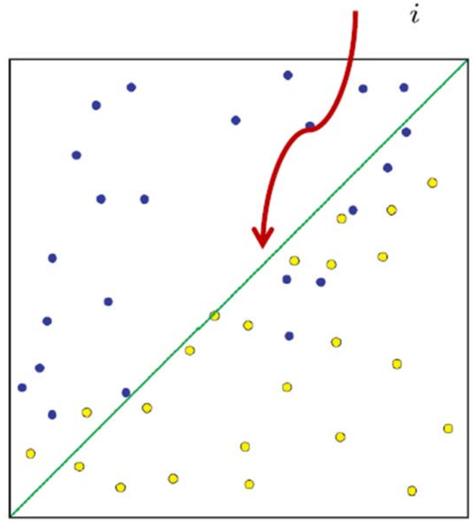


[Aarti, 10-701]

Estimating the parameters **w**

- Given $\{(X^{(1)}, Y^{(1)}), \ldots, (X^{(L)}, Y^{(L)})\}$
  - where $X^{(j)} = (X_1^{(j)}, X_2^{(j)}, \ldots, X_n^{(j)})$
- How to ***estimate*** **w** = (w_0, w_1, …, w_n)?

$$w_0 + \sum_i w_i X_i = 0$$



[Aarti, 10-701]

Estimating the parameters $\mathbf{w}$

- Given $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{L}$, $X^{(j)} = (X_1^{(j)}, X_2^{(j)}, \ldots, X_n^{(j)})$
- Assumptions: P(Y|X, w)

- Maximum ***conditional*** likelihood on data!

$$\widehat{\mathbf{w}}_{MCLE} = \arg\max_{\mathbf{w}} \prod_{j=1}^{L} P(Y^{(j)} \mid X^{(j)}, \mathbf{w})$$

  - Logistic regression only models P(Y|X)
  - So we only maximize P(Y|X), ***ignoring*** P(X)

Estimating the parameters $\mathbf{w}$

- Given $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{L}$, $X^{(j)} = (X_1^{(j)}, X_2^{(j)}, \ldots, X_n^{(j)})$
- Assumptions: $$P(Y=0|\mathbf{X},\mathbf{w}) = \frac{1}{1+exp(w_0+\sum_i w_i X_i)}$$ $$P(Y=1|\mathbf{X},\mathbf{w}) = \frac{exp(w_0+\sum_i w_i X_i)}{1+exp(w_0+\sum_i w_i X_i)}$$
- Maximum ***conditional*** likelihood on data!

$$\hat{\mathbf{w}}_{MCLE} = \arg\max_{\mathbf{w}} \prod_{j=1}^{L} P(Y^{(j)} \mid X^{(j)}, \mathbf{w})$$

  - Let's maximize conditional ***log***-likelihood

$$\max_{\mathbf{w}} \; l(\mathbf{w}) \equiv \ln \prod_{j}^{L} P(y^j|\mathbf{x}^j, \mathbf{w})$$

$$= \sum_{j}^{L} y^j (w_0 + \sum_{i}^{n} w_i x_i^j) - \ln(1 + exp(w_0 + \sum_{i}^{n} w_i x_i^j))$$

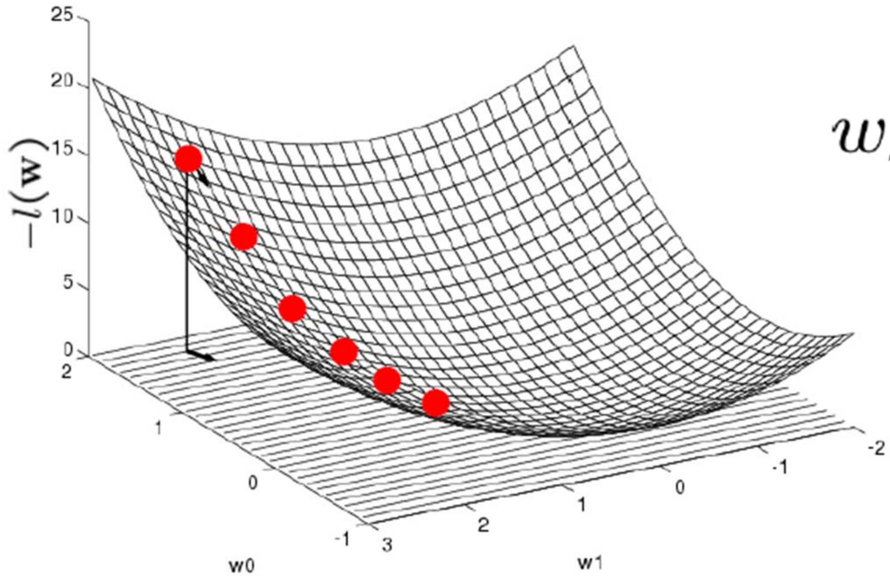
Estimating the parameters w

- Max conditional log-likelihood on data

$$\max_{\mathbf{w}} \; l(\mathbf{w}) \equiv \ln \prod_{j}^{L} P(y^j | \mathbf{x}^j, \mathbf{w})$$

$$= \sum_{j}^{L} y^j (w_0 + \sum_{i}^{n} w_i x_i^j) - \ln(1 + exp(w_0 + \sum_{i}^{n} w_i x_i^j))$$

  - A concave function (beyond the scope of class)
  - No local optimum: ***gradient ascent*** (descent) ☺



$$w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta \frac{\partial l(\mathbf{w})}{\partial w_i^{(t)}}$$

Estimating the parameters $\mathbf{w}$

- Max conditional log-likelihood on data

$$\max_{\mathbf{w}} \; l(\mathbf{w}) \equiv \ln \prod_{j}^{L} P(y^j|\mathbf{x}^j, \mathbf{w})$$

$$= \sum_{j}^{L} y^j (w_0 + \sum_{i}^{n} w_i x_i^j) - \ln(1 + exp(w_0 + \sum_{i}^{n} w_i x_i^j))$$

  - A concave function (beyond the scope of class)
  - No local optimum: ***gradient ascent*** (descent) ☺

$$w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta \frac{\partial l(\mathbf{w})}{\partial w_i^{(t)}}$$

$$w_i^{(t+1)} \leftarrow w_i^{(t)} + \eta \sum_{j=1}^{L} x_i^j [y^j - \hat{P}(Y^j = 1 \mid \mathbf{x}^j, \mathbf{w}^{(t)})]$$

Multiclass logistic regression

- Binary classification

$$P(Y=0|\mathbf{X},\mathbf{w}) = \frac{1}{1+exp(w_0+\sum_i w_i X_i)}$$

$$P(Y=1|\mathbf{X},\mathbf{w}) = \frac{exp(w_0+\sum_i w_i X_i)}{1+exp(w_0+\sum_i w_i X_i)}$$

- **K**-class classification
  - For each class k < K

$$P(Y=y_k|X) = \frac{\exp(w_{k0}+\sum_{i=1}^{d} w_{ki} X_i)}{1+\sum_{j=1}^{K-1}\exp(w_{j0}+\sum_{i=1}^{d} w_{ji} X_i)}$$

  - For class K

$$P(Y=y_K|X) = \frac{1}{1+\sum_{j=1}^{K-1}\exp(w_{j0}+\sum_{i=1}^{d} w_{ji} X_i)}$$

Outline

- Logistic regression
- Decision surface (boundary) of classifiers
  - Logistic regression
  - Gaussian naïve Bayes
  - Decision trees
- Generative vs. discriminative classifiers
- Linear regression
- Bias-variance decomposition and tradeoff
- Overfitting and regularization
- Feature selection

Logistic regression

- Model assumptions on P(Y|X):

$$P(Y=0|\mathbf{X},\mathbf{w}) = \frac{1}{1+exp(w_0+\sum_i w_i X_i)}$$

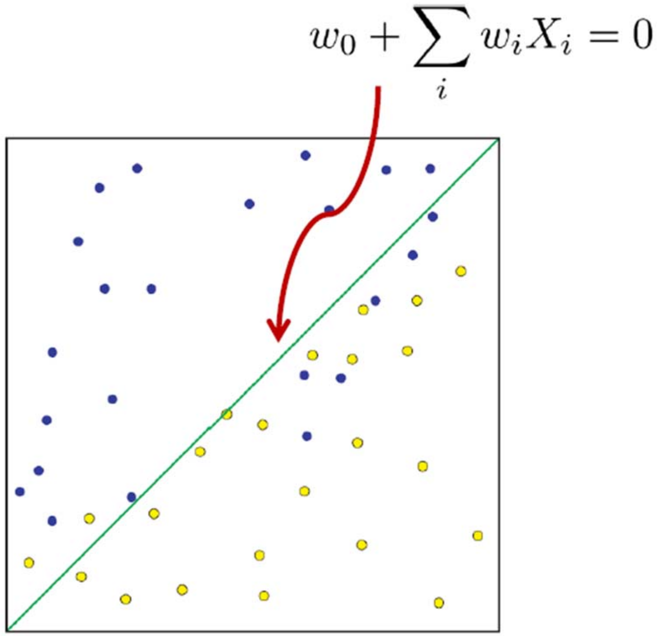
$$P(Y=1|\mathbf{X},\mathbf{w}) = \frac{exp(w_0+\sum_i w_i X_i)}{1+exp(w_0+\sum_i w_i X_i)}$$

- Deciding Y given X:

$$P(Y=0|X) \underset{1}{\overset{0}{\gtrless}} P(Y=1|X)$$

Linear decision boundary !

$$0 \underset{1}{\overset{0}{\gtrless}} w_0+\sum_i w_i X_i$$



[Aarti, 10-701]

Gaussian naïve Bayes

- Model assumptions P(X,Y) = P(Y)P(X|Y)
  - Bernoulli on Y: $P(Y=1)=\pi$
  - Conditional independence of X

    $$P(X=(X_1,X_2,...,X_n)|Y=k)=\prod_{i=1}^{n}P(X_i|Y=k)$$
  - Gaussian for X_i given Y: $P(X_i|Y=k) \sim N(\mu_{ik},\sigma_{ik}^2)$
- Deciding Y given X

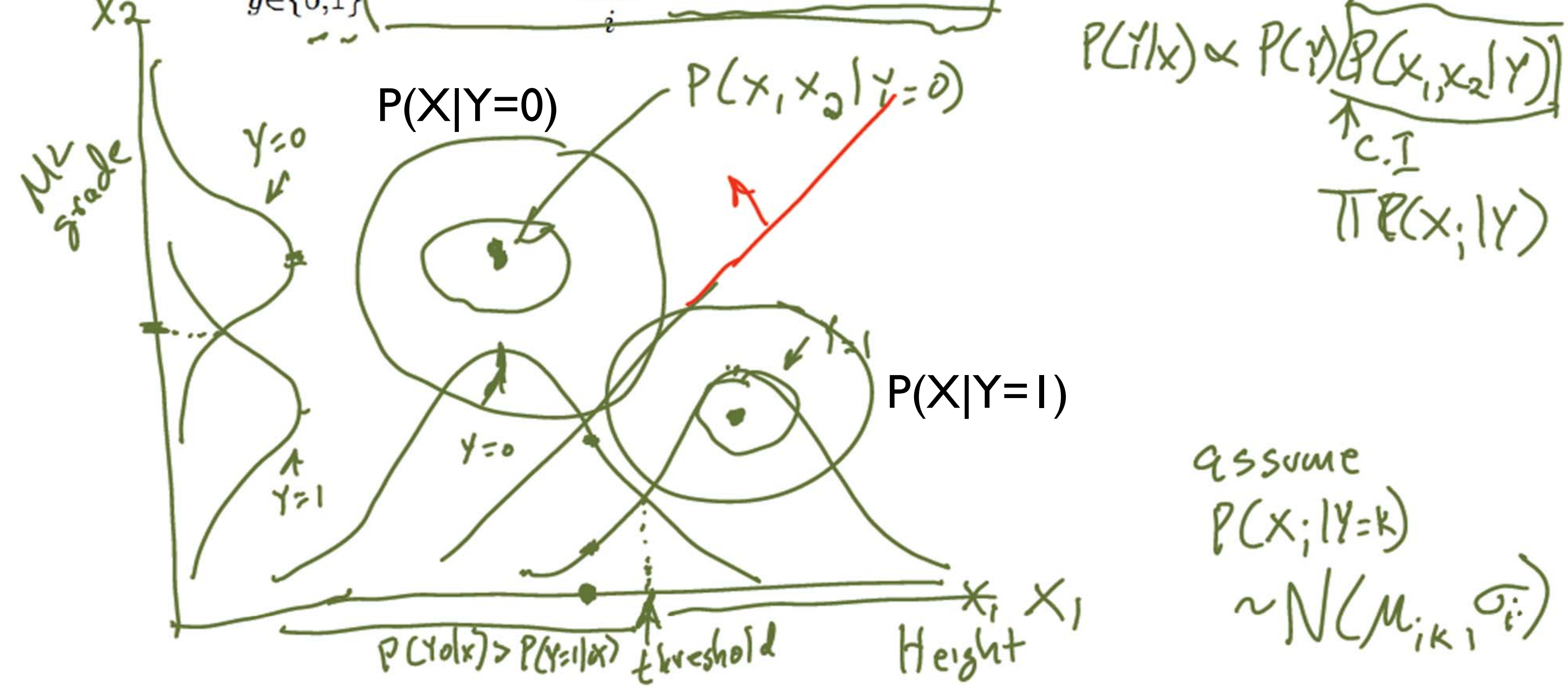
$$P(Y=0|X) \underset{1}{\overset{0}{\gtrless}} P(Y=1|X)$$

$$P(Y=0)P(X|Y=0) \underset{1}{\overset{0}{\gtrless}} P(Y=1)P(X|Y=1)$$

Gaussian Naïve Bayes – Big Picture

Consider boolean Y, continuous X_i. Assume P(Y=1)=0.5

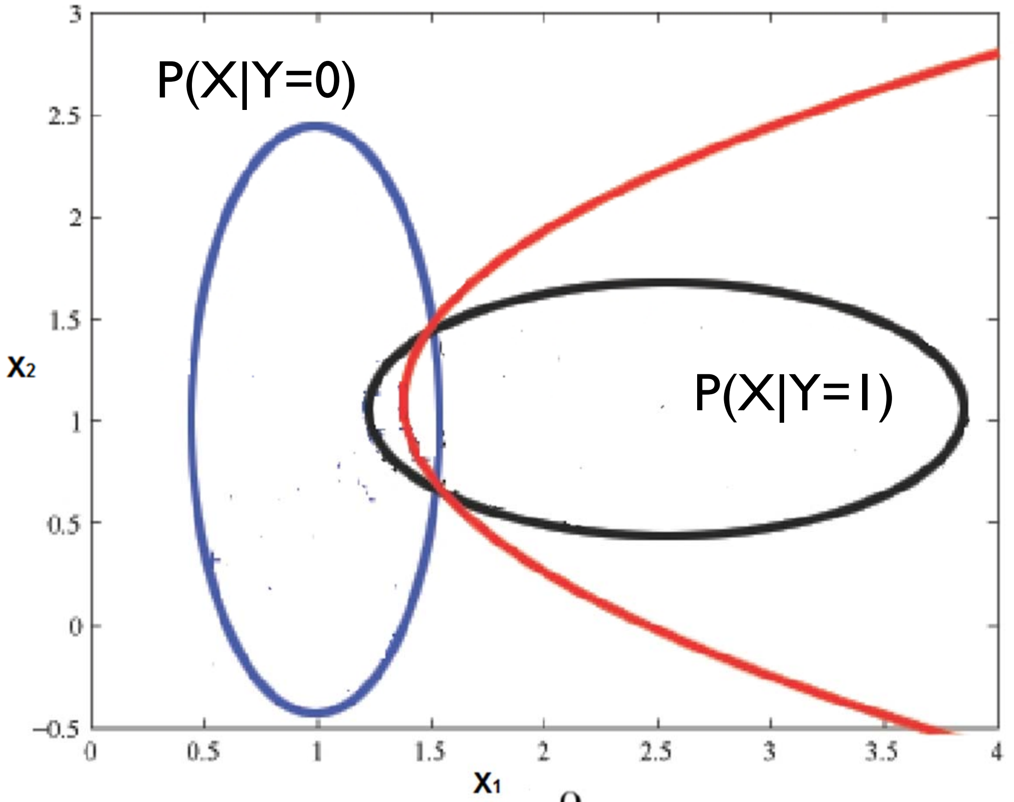
$$Y^{new} \leftarrow \arg\max_{y \in \{0,1\}} P(Y=y) \prod_i P(X_i^{new} | Y=y)$$



$$P(Y=0)P(X|Y=0) \underset{1}{\overset{0}{\gtrless}} P(Y=1)P(X|Y=1)$$

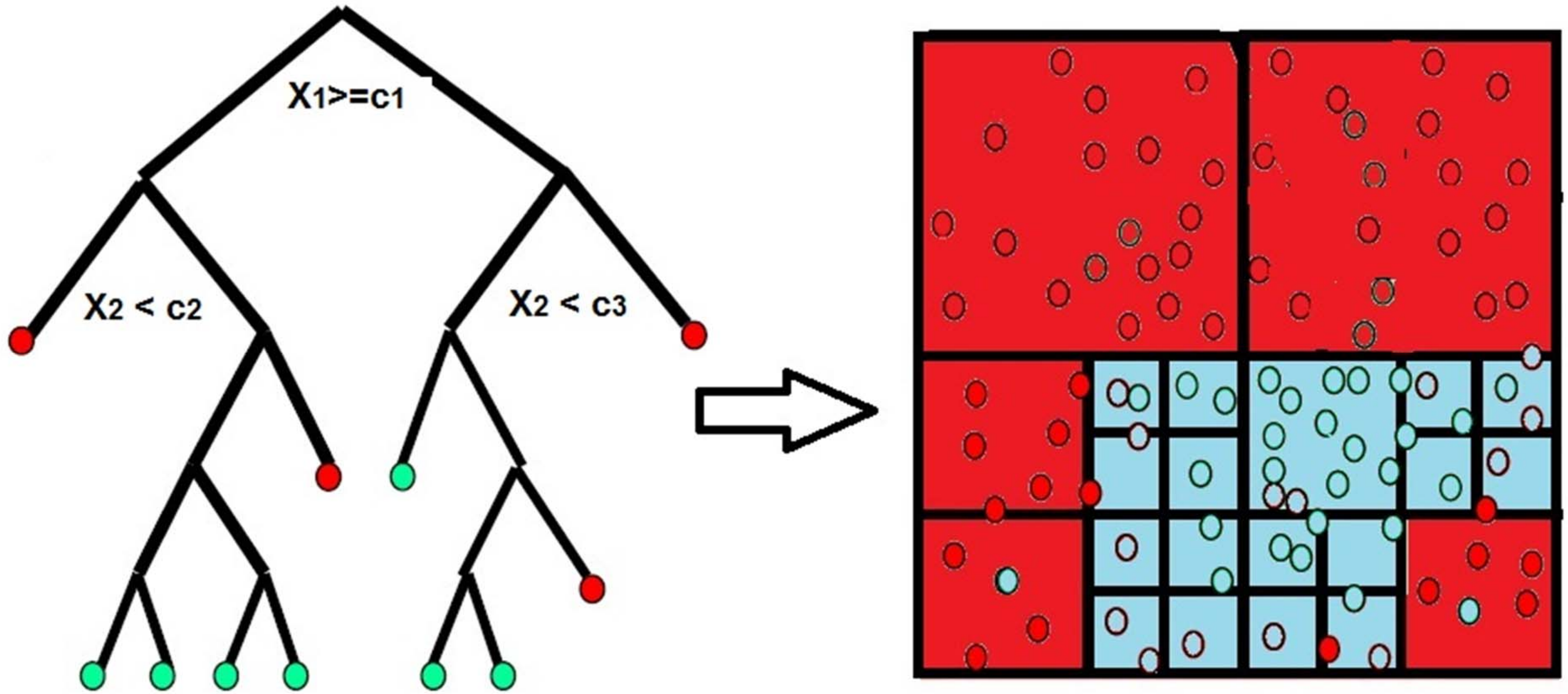
Gaussian naïve Bayes: *nonlinear* case

- Again, assume P(Y=1) = P(Y=0) = 0.5



$$P(Y=0)P(X|Y=0) \underset{1}{\overset{0}{\gtrless}} P(Y=1)P(X|Y=1)$$

Decision trees



- Decision making on Y: follow the tree structure to a leaf

Outline

- Logistic regression
- Decision surface (boundary) of classifiers
- Generative vs. discriminative classifiers
  - Definitions
  - How to compare them
  - GNB-1 vs. logistic regression
  - GNB-2 vs. logistic regression
- Linear regression
- Bias-variance decomposition and tradeoff
- Overfitting and regularization
- Feature selection

Generative and discriminative classifiers

- Generative classifiers
  - Modeling the joint distribution P(X,Y)
  - Usually via P(X,Y) = P(Y) P(X|Y)
  - Examples: Gaussian naïve Bayes ☺
- Discriminative classifiers
  - Modeling P(Y|X) **or** simply **f**: X→Y
  - Do not care about P(X)
  - Examples: logistic regression, support vector machines (later in this course)

Generative vs. discriminative

- How can we compare, say, Gaussian naïve Bayes and a logistic regression?
  - $P(X,Y) = P(Y)\ P(X|Y)$ vs. $P(Y|X)$?
- Hint: decision making is based on $P(Y|X)$
  - Compare the $P(Y|X)$ they can represent !

Two versions: GNB-1 and GNB-2

- Model assumptions on P(X,Y) = P(Y)P(X|Y)
  - Bernoulli on Y: $P(Y=1)=\pi$
  - Conditional independence of X

    $$P(X=(X_1,X_2,...,X_n)|Y=k)=\prod_{i=1}^{n}P(X_i|Y=k)$$

GNB-1
  - Gaussian on X_i|Y: $P(X_i|Y=k) \;\sim\; N(\mu_{ik},\underline{\sigma_{ik}^2})$

GNB-2
  - (Additionally,) class-independent variance

    $$P(X_i|Y=k) \;\sim\; N(\mu_{ik},\underline{\sigma_i^2})$$

Two versions: GNB-1 and GNB-2

- Model assumptions on P(X,Y) = P(Y)P(X|Y)
  - Bernoulli on Y: $P(Y=1)=\pi$
  - Conditional independence of X

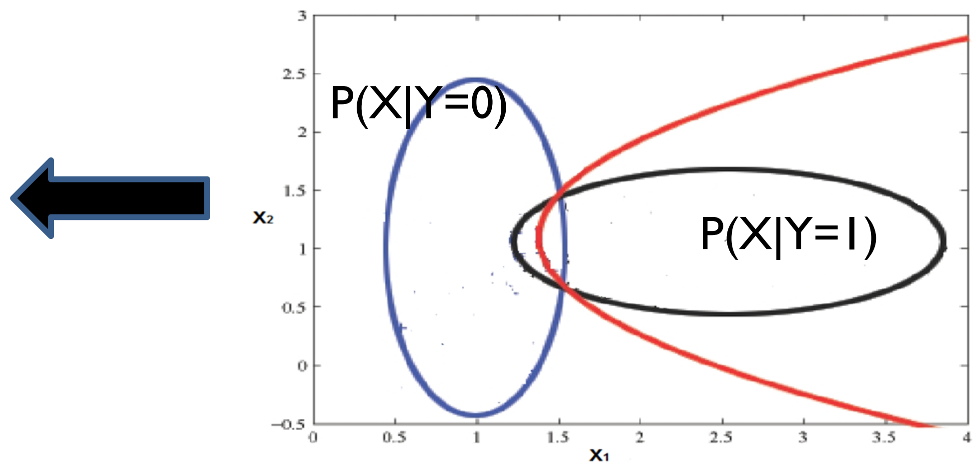
    $$P(X=(X_1,X_2,\ldots,X_n)|Y=k)=\prod_{i=1}^{n}P(X_i|Y=k)$$

GNB-1 ◦ Gaussian on $X_i|Y$: $P(X_i|Y=k)\ \sim\ N(\mu_{ik},\underline{\sigma^2_{ik}})$

GNB-2 ◦ (Additionally,) class-independent variance

$$P(X_i|Y=k)\ \sim\ N(\mu_{ik},\underline{\sigma^2_i})$$

Impossible for GNB-2



GNB-2 vs. logistic regression

- GNB-2: P(X,Y) = P(Y)P(X|Y)
  - Bernoulli on Y: $P(Y=1)=\pi$
  - Conditional independence of X, and Gaussian on X_i
  - Additionally, class-independent variance

$$P(X_i|Y=k) \ \sim \ N(\mu_{ik}, \underline{\sigma_i^2})$$

- It turns out, P(Y|X) of GNB-2 has the form:

$$P(Y=1|X) = \frac{1}{1+\exp(\ln\frac{1-\pi}{\pi}+\sum_i\left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}X_i+\frac{\mu_{i1}^2-\mu_{i0}^2)}{2\sigma_i^2}\right))}$$

GNB-2 vs. logistic regression

- It turns out, P(Y|X) of GNB-2 has the form:

$$P(Y=1|X) = \frac{1}{1+\exp(\ln\frac{1-\pi}{\pi} + \sum_i \left(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}X_i + \frac{\mu_{i1}^2-\mu_{i0}^2)}{2\sigma_i^2}\right))}$$

  - See [Mitchell: Naïve Bayes and Logistic Regression], section 3.1 (page 8 – 10)

- Recall: P(Y|X) of logistic regression:

$$P(Y=1|X) = \frac{1}{1+\exp(w_0 + \sum_{i=1}^n w_i X_i)}$$

GNB-2 vs. logistic regression

- P(Y|X) of GNB-2 is ***subset*** of P(Y|X) of LR
- Given *infinite* training data
  - We claim: LR >= GNB-2

GNB-1 vs. logistic regression

- GNB-1: P(X,Y) = P(Y)P(X|Y)
  - Bernoulli on Y: $P(Y=1)=\pi$
  - Conditional independence of X, and Gaussian on X_i

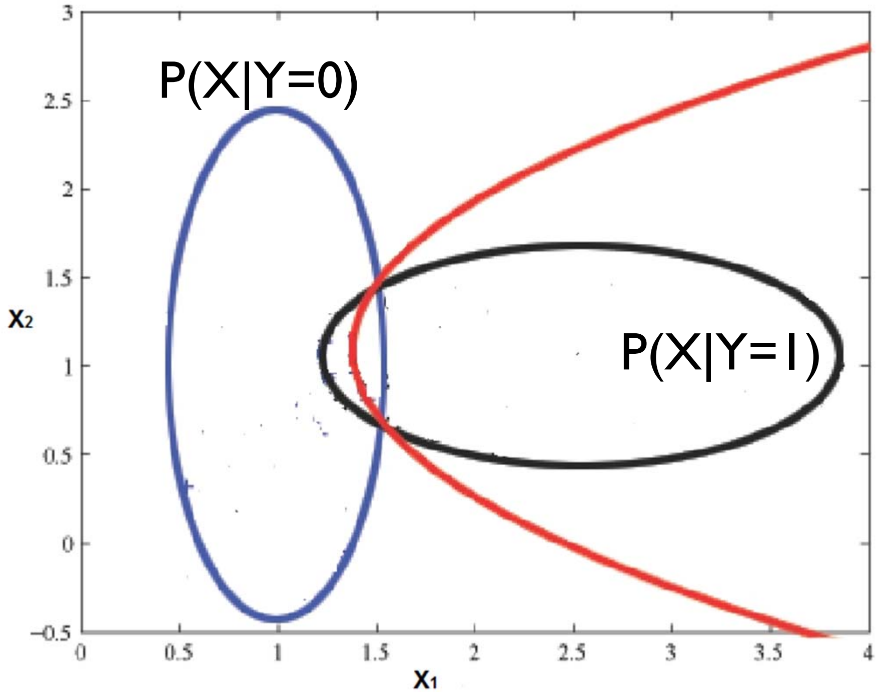
$$P(X_i|Y=k) \sim N(\mu_{ik}, \underline{\sigma_{ik}^2})$$

- Logistic regression: P(Y|X)

$$P(Y=1|X) = \frac{1}{1+\exp(w_0+\sum_{i=1}^{n} w_i X_i)}$$

GNB-1 vs. logistic regression

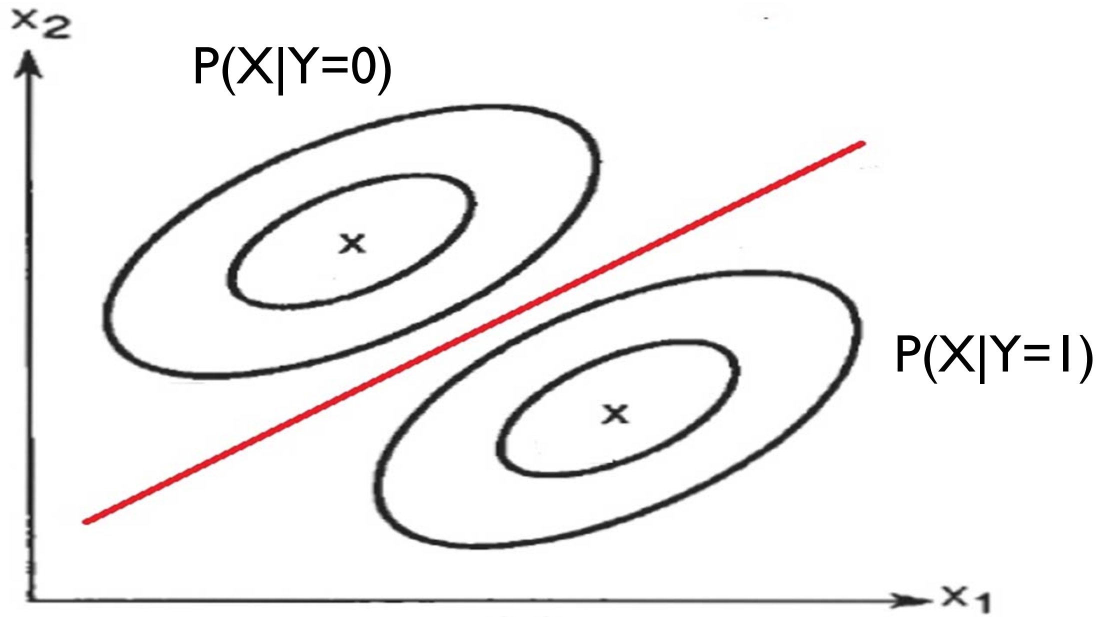
- ***None of them*** encompasses the other
- First, find a P(Y|X) from GNB-1 that ***cannot*** be represented by LR



  - LR only represents linear decision surfaces

GNB-1 vs. logistic regression

- ***None of them*** encompasses the other
- Second, find a P(Y|X) represented by LR that ***cannot*** be derived from GNB-1assumptions



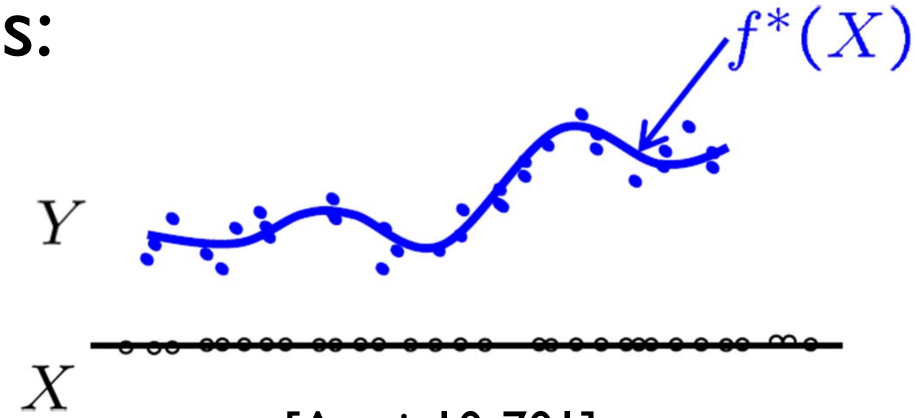
  - GNB-1 cannot represent any correlated Gaussian
  - But can still possibly be represented by LR (HW2)

Outline

- Logistic regression
- Decision surface (boundary) of classifiers
- Generative vs. discriminative classifiers
- Linear regression
  - Regression problems
  - Model assumptions: P(Y|X)
  - Estimate the model parameters
- Bias-variance decomposition and tradeoff
- Overfitting and regularization
- Feature selection

Regression problems

- Regression problems:
  - Predict Y given X
  - Y is continuous



[Aarti, 10-701]

  - General assumption:

$$Y = f^*(X) + \epsilon, \qquad \epsilon \sim Distribution()$$

Linear regression: assumptions

- Linear regression assumptions
  - Y is generated from *f*(X) plus Gaussian noise

$$Y = f(X) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

  - *f*(X) is a linear function

$$f(X) = w_0 + \sum_{i=1}^{n} w_i X_i$$

Linear regression: assumptions

- Linear regression assumptions
  - Y is generated from *f*(X) plus Gaussian noise

$$Y = f(X) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

  - *f*(X) is a linear function

$$f(X) = w_0 + \sum_{i=1}^{n} w_i X_i$$

- Therefore, assumptions on P(Y|X, w):

$$\begin{aligned} P(Y|X = (X_1, X_2, \ldots, X_n)) &\sim N(f(X), \sigma^2) \\ &\sim N(w_0 + \textstyle\sum_{i=1}^{n} w_i X_i,\ \sigma^2) \end{aligned}$$

Linear regression: assumptions

- Linear regression assumptions
  - Y is generated from *f*(X) plus Gaussian noise

$$Y = f(X) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

  - *f*(X) is a linear function

$$f(X) = w_0 + \sum_{i=1}^{n} w_i X_i$$

- Therefore, assumptions on P(Y|X, w):

$$P(Y|X = (X_1, X_2, \ldots, X_n)) \sim N(f(X), \sigma^2)$$

$$\sim N(w_0 + \sum_{i=1}^{n} w_i X_i, \ \sigma^2)$$

$$P(Y|X, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(Y - w_0 - \sum_{i=1}^{n} w_i X_i)^2}{2\sigma^2}$$

Estimating the parameters **w**

- Given $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{L}$, $X^{(j)} = (X_1^{(j)}, X_2^{(j)}, \ldots, X_n^{(j)})$
- Assumptions:

$$P(Y|X, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(Y - w_0 - \sum_{i=1}^{n} w_i X_i)^2}{2\sigma^2}$$

- Maximum ***conditional*** likelihood on data!

$$\widehat{\mathbf{w}}_{MCLE} = \arg\max_{\mathbf{w}} \prod_{j=1}^{L} P(Y^{(j)} \mid X^{(j)}, \mathbf{w})$$

Estimating the parameters $\mathbf{w}$

- Given $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{L}$, $X^{(j)} = (X_1^{(j)}, X_2^{(j)}, \ldots, X_n^{(j)})$
- Assumptions:

$$P(Y|X, \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(Y - w_0 - \sum_{i=1}^{n} w_i X_i)^2}{2\sigma^2}$$

- Maximum ***conditional*** likelihood on data!

$$\widehat{\mathbf{w}}_{MCLE} = \arg\max_{\mathbf{w}} \prod_{j=1}^{L} P(Y^{(j)} \mid X^{(j)}, \mathbf{w})$$

  - Let's maximize conditional ***log***-likelihood

Estimating the parameters w

- Given $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{L}$, $X^{(j)} = (X_1^{(j)}, X_2^{(j)}, \ldots, X_n^{(j)})$
- Assumptions:

$$P(Y|X,\mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(Y-w_0-\sum_{i=1}^{n} w_i X_i)^2}{2\sigma^2}$$

- Maximum ***conditional*** likelihood on data!

$$\widehat{\mathbf{w}}_{MCLE} = \arg\max_{\mathbf{w}} \prod_{j=1}^{L} P(Y^{(j)} \mid X^{(j)}, \mathbf{w})$$

  - Let's maximize conditional ***log***-likelihood

$$\max_{\mathbf{w}} \; l(\mathbf{w}) \equiv \ln \prod_{j}^{L} P(y^j|\mathbf{x}^j, \mathbf{w})$$

$$= \sum_{j=1}^{L} -\frac{(Y^{(j)}-w_0-\sum_{i=1}^{n} w_i X_i^{(j)})^2}{2\sigma^2} + const$$

Estimating the parameters $\mathbf{w}$

- Max the conditional log-likelihood over data

$$\arg\max_{\mathbf{w}} \sum_{j=1}^{L} -\frac{(Y^{(j)}-w_0-\sum_{i=1}^{n} w_i X_i^{(j)})^2}{2\sigma^2}$$

$$= \arg\max_{\mathbf{w}} \sum_{j=1}^{L} -(Y^{(j)} - w_0 - \sum_{i=1}^{n} w_i X_i^{(j)})^2$$

  - OR minimize the sum of "squared errors"

$$= \arg\min_{\mathbf{w}} \sum_{j=1}^{L} (Y^{(j)} - w_0 - \sum_{i=1}^{n} w_i X_i^{(j)})^2$$

  - Gradient ascent (descent) is easy
  - Actually, a closed form solution exists ☺

Estimating the parameters **w**

- Max the conditional log-likelihood over data

$$\text{argmax}_{\mathbf{w}} \sum_{j=1}^{L} -(Y^{(j)} - w_0 - \sum_{i=1}^{n} w_i X_i^{(j)})^2$$

  - OR

$$\text{argmin}_{\mathbf{w}} \sum_{j=1}^{L} (Y^{(j)} - w_0 - \sum_{i=1}^{n} w_i X_i^{(j)})^2$$

- Actually, a closed form solution exists ☺
  - w = $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y}$
  - **A** is an *L* by *n* matrix: m examples, n variables
  - ***Y*** is an *L* by *1* vector: m examples

Outline

- Logistic regression
- Decision surface (boundary) of classifiers
- Generative vs. discriminative classifiers
- Linear regression
- Bias-variance decomposition and tradeoff
  - True risk for a (regression) model
  - Risk of the perfect model
  - Risk of a learning method: bias and variance
  - Bias-variance tradeoff
- Overfitting and regularization
- Feature selection

(True) risk

- Consider a regression problem
- (True) risk of a prediction model f( )

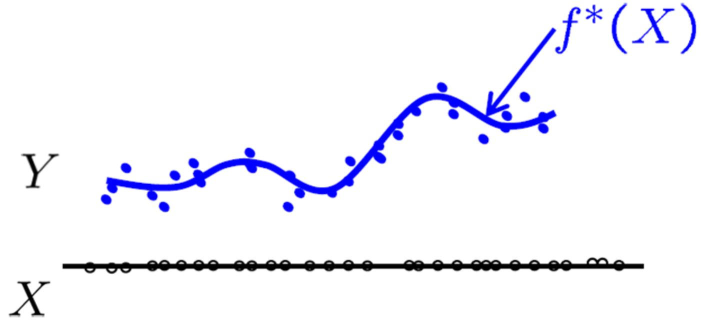
$$R(f) = E_{X,Y}[(f(X) - Y)^2]$$

  - ***Expected*** squared error when we use f( ) to make prediction on future examples

Risk of the perfect model

- The "true" model f*( )

$$Y = f^*(X) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$



[Aarti, 10-701]

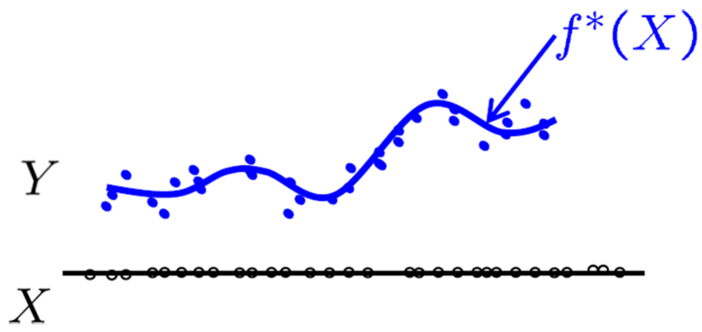
- The risk of f*( )

$$R^* = \mathbb{E}_{XY}[(f^*(X) - Y)^2] = \mathbb{E}[\epsilon^2] = \sigma^2$$

  - The best we can do !
  - σ^2 is the "unavoidable risk"
- Is this achievable ? … well …
  - Model makes perfect assumptions
    - f*() belongs to the class of functions we consider
  - Infinite training data

A learning method

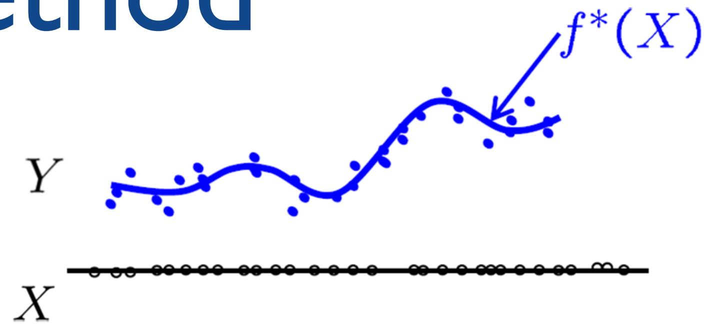
Regression: $Y = f^*(X) + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$



- A learning method:
  - Model assumptions, i.e., the space of models
    - e.g., the form of P(Y|X) in linear regression
  - An optimization/search algorithm
    - e.g., maximum conditional likelihood on data
- Given a set of *L* training samples
  - ***D*** = $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{L}$
  - A learning method outputs: $\hat{f}_D(\ )$

Risk of a learning method

Regression: $Y = f^*(X) + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$

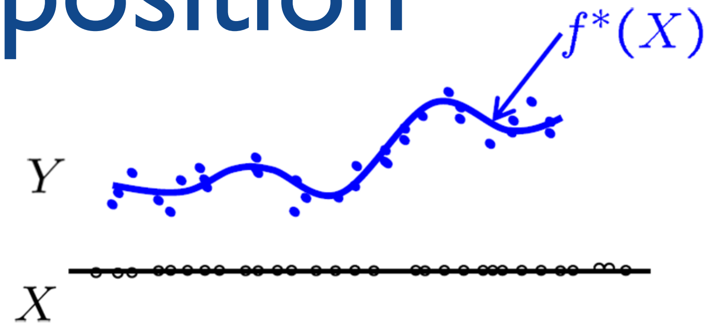


- Risk of a learning method

$$E_D[R(\hat{f}_D)] = E_{X,Y,D}[(\hat{f}_D(X) - Y)^2]$$

Bias-variance decomposition

Regression: $Y = f^*(X) + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$



- Risk of a learning method

$$E_D[R(\hat{f}_D)] = E_{X,Y,D}[(\hat{f}_D(X) - Y)^2]$$

$$= \quad E_{X,Y}[(\underline{E_D[\hat{f}_D(X)] - f^*(X))^2}]$$

Bias^2

$$+$$

$$E_{X,Y}[\ \underline{E_D[(\hat{f}_D(X) - E_D[\hat{f}_D(X)])^2]}\ ]$$

Variance

$$+$$

$$\underline{\sigma^2}$$

Unavoidable error

bias : $E[Z] - Z^*$

variance : $E[(Z - E[Z])^2]$

Bias-variance decomposition

$$E_D[R(\hat{f}_D)] = E_{X,Y,D}[(\hat{f}_D(X) - Y)^2]$$

$$= \quad E_{X,Y}[\underline{(E_D[\hat{f}_D(X)] - f^*(X))^2}] \quad \text{Bias}^2$$

$$+$$

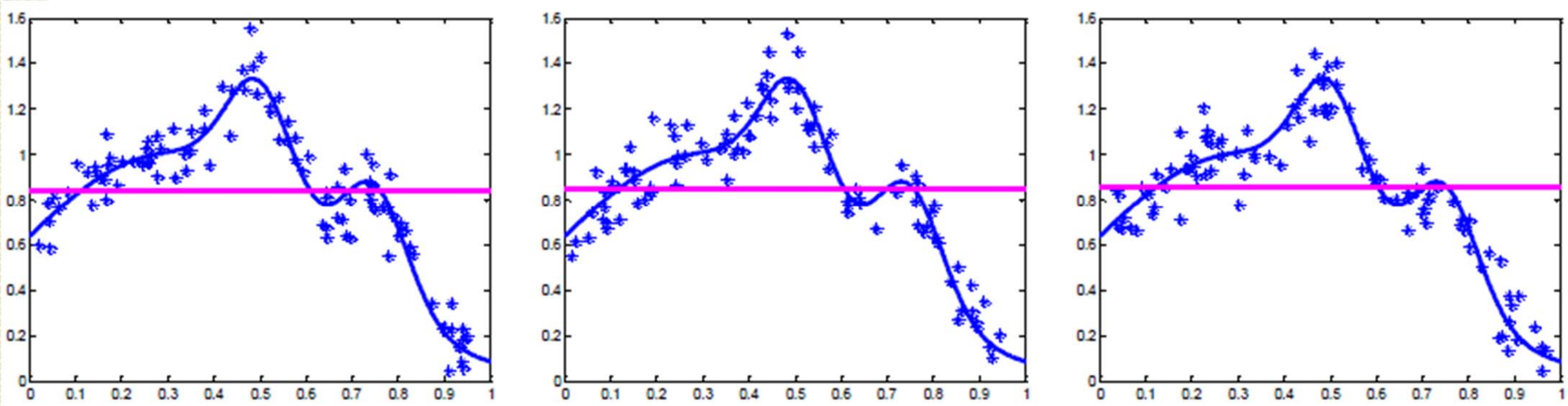
$$E_{X,Y}[\ \underline{E_D[(\hat{f}_D(X) - E_D[\hat{f}_D(X)])^2]}\ ] \quad \text{Variance}$$

$$+ \ \sigma^2$$

- ***Bias***: how much is the "mean" estimation different from the true function f*
  - Does f* belong to our model space? If not, how far?
- ***Variance***: how much is a single estimation different from the "mean" estimation
  - How sensitive is our method to a "bad" training set ***D*** ?

Bias-variance tradeoff

- Minimize both bias and variance ? No free lunch
- Simple models: low variance but high bias

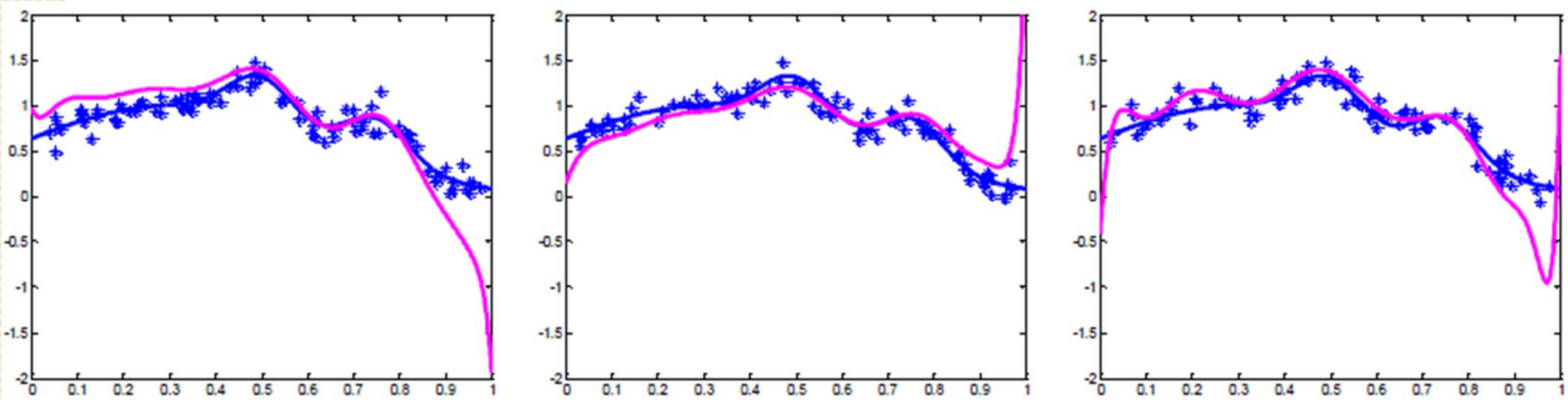


[Aarti]

  - Results from 3 random training sets ***D***
  - Estimation is very stable over 3 runs (low variance)
  - But estimated models are ***too simple*** (high bias)

Bias-variance tradeoff

- Minimize both bias and variance ? No free lunch
- Complex models: low bias but high variance



  - Results from 3 random training sets ***D***
  - Estimated models complex enough (low bias)
  - But estimation is ***unstable*** over 3 runs (high variance)

Bias-variance tradeoff

- We need a good tradeoff between bias and variance
  - Class of models are not too simple (so that we can ***approximate*** the true function well)
  - But not too complex to overfit the training samples (so that the ***estimation*** is ***stable***)

Outline

- Logistic regression
- Decision surface (boundary) of classifiers
- Generative vs. discriminative classifiers
- Linear regression
- Bias-variance decomposition and tradeoff
- Overfitting and regularization
  - Empirical risk minimization
  - Overfitting
  - Regularization
- Feature selection

Empirical risk minimization

- Many learning methods essentially minimize a loss (risk) function over training data

$$\text{argmin}_{\mathbf{w}} \sum_{j=1}^{L} \underline{R(Y^{(j)}, X^{(j)}, \mathbf{w})}$$

  - Both linear regression and logistic regression

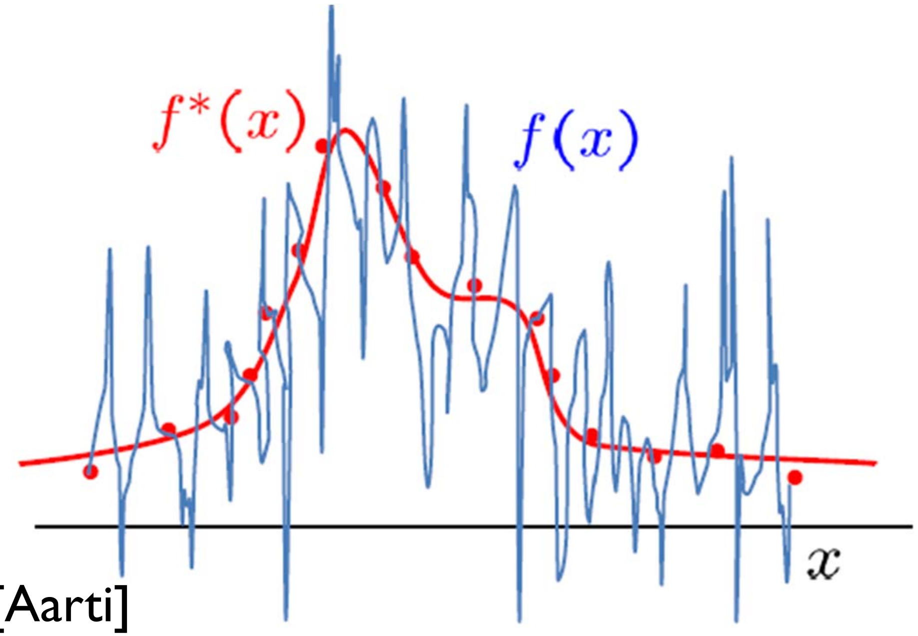
$$\begin{aligned} &\text{argmax}_{\mathbf{w}} \ln \prod_{j=1}^{L} P(Y^{(j)}|X^{(j)}, \mathbf{w}) \\ &= \text{argmin}_{\mathbf{w}} - \ln \prod_{j=1}^{L} P(Y^{(j)}|X^{(j)}, \mathbf{w}) \\ &= \text{argmin}_{\mathbf{w}} \sum_{j=1}^{L} \underline{- \ln P(Y^{(j)}|X^{(j)}, \mathbf{w})} \end{aligned}$$

  - (linear regression: squared errors)

$$= \text{argmin}_{\mathbf{w}} \sum_{j=1}^{L} \underline{(Y^{(j)} - w_0 - \sum_{i=1}^{n} w_i X_i^{(j)})^2}$$

Overfitting

- Minimize the loss over ***finite*** samples is not always a good thing … overfitting!

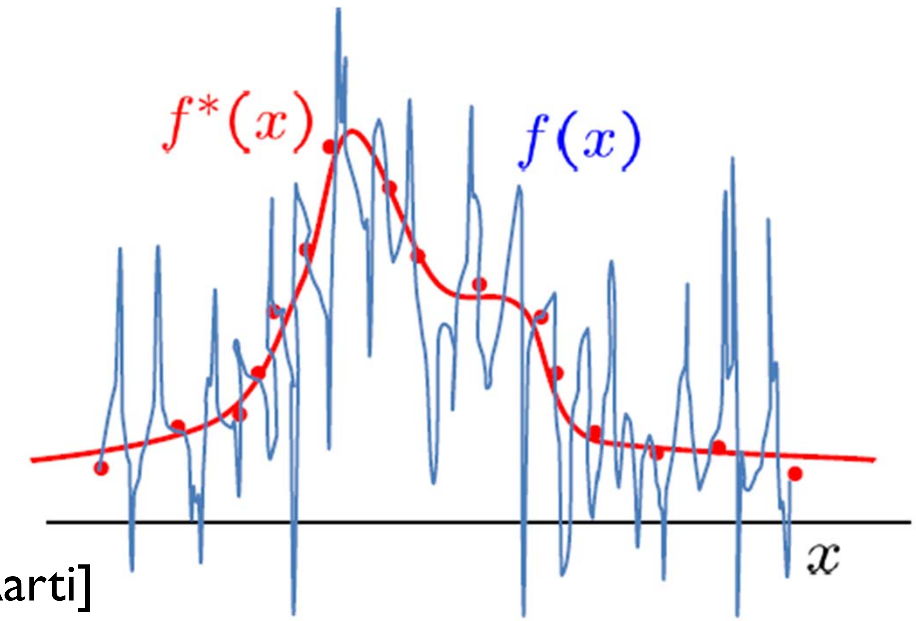


[Aarti]

  - The blue function has ***zero*** error on training samples, but is ***too complicated*** (and “crazy”)

Overfitting

- How to avoid overfitting?



[Aarti]

  - Hint: complicated and "crazy" models like the blue one usually have ***large*** model coefficients ***w***

Regularization: L-2 regularization

- Minimize the loss + a regularization penalty

$$\text{argmin}_{\mathbf{w}} - \ln \prod_{j=1}^{L} P(Y^{(j)}|X^{(j)}) \; + \; \underline{\lambda} \, \underline{\sum_{i=1}^{n} w_i^2}$$

- Intuition: prefer ***small*** coefficients
- Prob. interpretation: a Gaussian prior on **w**
  - Minimize the penalty $\lambda \sum_{i=1}^{n} w_i^2$ is:

$$min_{\mathbf{w}} - \ln \prod_{i=1}^{n} p(w_i), \quad p(w_i) \sim N(0, 1/\lambda)$$

  - So minimize loss+penalty is max (log-)posterior

$$\text{argmin}_{\mathbf{w}} - \ln \prod_{j=1}^{L} P(Y^{(j)}|X^{(j)}) \; - \; \ln \prod_{i=1}^{n} p(w_i)$$

$$\text{argmax}_{\mathbf{w}} \underline{\ln \prod_{j=1}^{L} P(Y^{(j)}|X^{(j)})} \; + \; \underline{\ln \prod_{i=1}^{n} p(w_i)}$$

Regularization: L-1 regularization

- Minimize a loss + a regularization penalty

$$\arg\min_{\mathbf{w}} - \ln \prod_{j=1}^{L} P(Y^{(j)}|X^{(j)}) \;+\; \underline{\lambda \sum_{i=1}^{n} |w_i|}$$

- Intuition: prefer ***small*** and ***sparse*** coefficients
- Prob. interpretation: a Laplacian prior on **w**

Outline

- Logistic regression
- Decision surface (boundary) of classifiers
- Generative vs. discriminative classifiers
- Linear regression
- Bias-variance decomposition and tradeoff
- Overfitting and regularization
- Feature selection

Feature selection

- Feature selection: select a subset of "useful" features
  - Less features → less complex models
  - Lower "variance" in the risk of the learning method ☺
  - But might lead to higher "bias" ☹

Feature selection

- ***Score each feature*** and select a subset
  - The mutual information between X_i and Y

$$\hat{I}(X_i, Y) = \sum_k \sum_y \hat{P}(X_i = k, Y = y) \log \frac{\hat{P}(X_i = k, Y = y)}{\hat{P}(X_i = k)\hat{P}(Y = y)}$$

  - Accuracy of single-feature classifier f: $X_i \rightarrow Y$
  - etc.

Feature selection

- Score each feature and ***select a subset***
  - One-step method: select *k* highest score features
  - Iterative method:
    - Select a highest score feature from the pool
    - ***Re-score*** the rest, e.g., mutual information ***conditioned*** on already-selected features
    - Iterate

Outline

- Logistic regression
- Decision surface (boundary) of classifiers
- Generative vs. discriminative classifiers
- Linear regression
- Bias-variance decomposition and tradeoff
- Overfitting and regularization
- Feature selection

Homework 2 due tomorrow

- Feb. 4th (Friday), 4pm
- Sharon Cavlovich's office (GHC 8215).
- 3 separate sets (each question for a TA)

The last slide

- **Go Steelers !**

- Question ?