CS 307: Lecture 17

$P(x|y)$, $P(y)$

<u>Maximum Likelihood Estimation</u>

Find parameters $P(x|y)$ and $P(y)$ that maximize the probability of observing the training set.

<u>MAP rule</u>: Once parameters are discovered, on a new example (x)

$$p(y=0|x) = \frac{p(x|y=0)\,p(y=0)}{p(x|y=0)\cdot p(y=0) + p(x|y=1)\cdot p(y=1)}$$

$$p(y=1|x) = \frac{p(x|y=1)\,p(y=1)}{p(x)}$$

Output 1 if $p(y=1|x) > p(y=0|x)$

0 otherwise

<u>Gaussian Discriminant Analysis</u>

$$p(x|y) \sim N(\mu_y, \Sigma) \leftarrow$$

$$p(y) \sim Ber(\phi)$$

MAP rule: $\log \left[\frac{p(y=1|x)}{p(y=0|x)} \right] > 0$

$\left[p(y=0|x) \right]$

$\Rightarrow$ linear classifier.

$$p(y=1|x) = \frac{1}{1+e^{-\theta^T x}}$$

Naive Bayes:

Classification: Given an email classify if it is spam.

Input features: Given emal

$$\begin{bmatrix} \\ 0 \\ \\ \\ \end{bmatrix}$$

$\leftarrow$ One position for every word in Some vocabulary

0 if this word does not appear in the email

1 if this word appears in email.

Discriminative Learning Algorithm

$p(y=1|x) \leftarrow$ Learn

$\hookrightarrow \mathrm{Ber}(\phi_x)$

parameters is $\underline{2^d}$ $(x \in \mathbb{R}^d)$

Generative Learning Algorithm

parameters : $p(y=1) = \phi_y$

$p(x \mid y=0)$

$p(x \mid y=1)$

Naive Bayes assumption :

Probability that the j th word appears in an email (spam/non-spam) is independent of the k th word appearing in an email.

$$p(x \mid y=0) = \prod_{j=1}^{d} p(x_j = 1 \mid y=0)$$

$$p(x_j = 1 \mid y=0) = \phi_{j|0}$$

$$p(x_j = 1 \mid y=1) = \phi_{j|1}$$

— 2d

parameters : $p(y=1) = \phi_y$

If $y = i \in \{0,1\}$ $p(x_j = 1 \mid y=i) = \phi_{j|i}$

$$L(\phi_y, \phi_{j|0}, \phi_{j|1})$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{d} \phi_{j|y}^{1[x_j^{(i)}=1]} (1-\phi_{j|y})^{1[x_j^{(i)}=0]}$$

$$\phi^{1[y=1]} (1-\phi)^{1[y=0]}$$

$\phi_y \quad (1-\phi_y)$

We will maximize the log of the likelihood :

$$\phi_y = \frac{\sum_{i=1}^{n} 1[y^{(i)}=1]}{}$$

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{n} 1[x_j^{(i)}=1 \wedge y^{(i)}=0]}{\sum_{i=1}^{n} 1[y^{(i)}=0]}$$

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{n} 1[x_j^{(i)}=1 \wedge y^{(i)}=1]}{\sum_{i=1}^{n} 1[y^{(i)}=1]}$$

$$\phi_{j=\text{COVID}|0} = 0$$

$$\phi_{j=\text{COVID}|1} = 0$$

$$P(y=1|x) = \frac{p(x|y=1)\cdot p(y=1)}{\underbrace{p(x|y=0)\cdot p(y=0)}_{0} + \underbrace{p(x|y=1)\cdot p(y=1)}_{0}}$$

$$= \frac{\prod_{j=1}^{d} \phi_{j|1}\, \phi_y}{\cdots\cdots}$$

$$= \frac{0}{0}$$

Laplace Smoothing :

Laplace Smoothing:

Random Variable $\underbrace{Z \sim \text{Multinomial}(k)}_{Z \in \{1,2..k\},\ \phi_i = P[Z=i]}$

$$\phi_j = \frac{1 + \sum_{i=1} 1[z=j]}{k+n}$$

Back to email classification

$$\phi_{j|y=s} = \frac{1 + \sum 1[x_j^{(i)}=1 \wedge y^{(i)}=s]}{2 + \sum 1[y^{(i)}=s]}$$

Generalization:

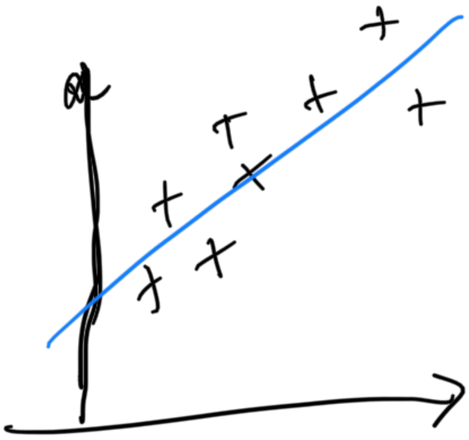
Goal of learning: Discover a model of the world that explains data.

Minimizing training is a way to try to ensure that our learnt model will do well on new test data.

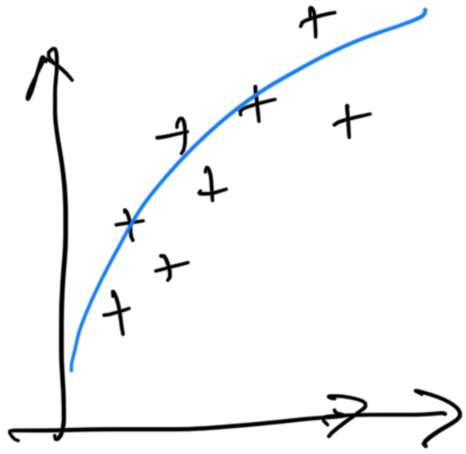
Assuming: There is distribution D and traing set is generated by sampling from D and new examples will come from D.

Real goal is minimize

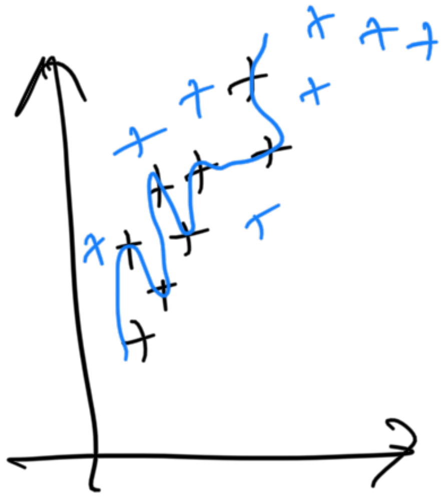
$$E_{x \sim D}[\text{loss}]$$



Linear



Quadratic



High poly

Linear
High traing error
Regression

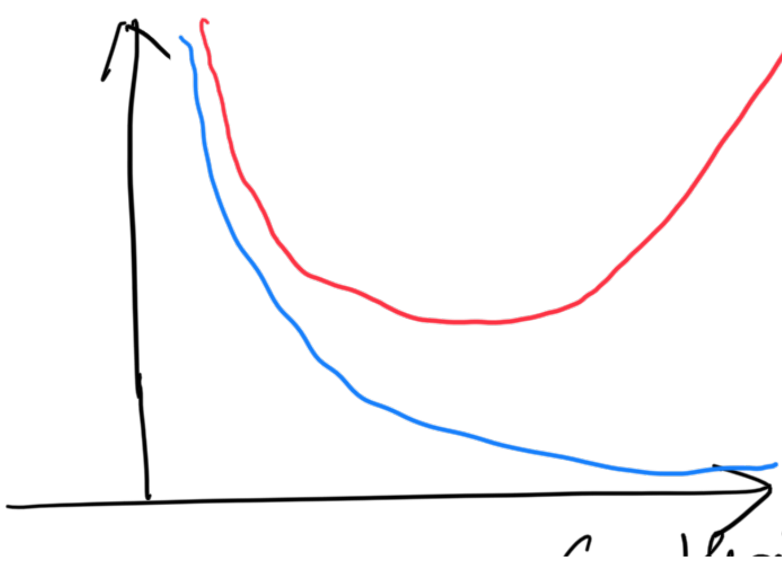
High Bias
Underfitting — High traing + test error

Cannot addressed by more traing data

High poly
Lowest traing error
High Variance
Overfitting.
Low traing error
High test error

Addressed increasing traing data.

Bias / Variance



Complexity of Model