

Three Topics:

(I) Receiver Operator Characteristic (ROC) curves
Precision and Recall Curves.

* What if we do not fix the loss function?

(II) The Curse of Dimensionality

* Why human intuitions about geometry are misleading in high-dimensional spaces.

* Why we do not have enough data in high-dimensions. And what we can do about it

III The Bias-Variance Dilemma.

* A classic statistical perspective about generalization.

Spring 2013

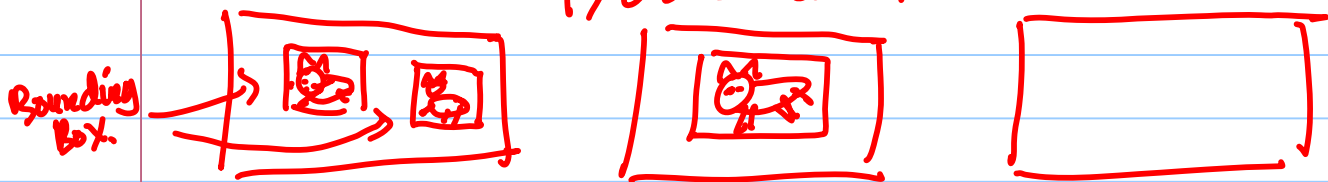
(2)

Topic I: Precision/Recall and ROC curve.

What if we do not know the loss function?
Or if we want a more "sophisticated" decision process. i.e. we don't want to diagnose people as "cancer" or "non-cancer". Instead we want to separate them into groups for further testing.

Example: Detecting cats in the Pascal Challenge

Data: 20,000 images
1,000 cats.



Images can contain 0 cats, 1 cat, or more cats.

The position of cats are specified by bounding boxes surrounding them

Task: determine the boxes which contain cats

• There are 1,000 boxes that contain cats

• There are $20,000 \times 1,000 = 20,000,000$

$\underbrace{\hspace{2em}}_{\text{no. images}} \underbrace{\hspace{2em}}_{\substack{\text{no. of boxes} \\ \text{in image}}} \underbrace{\hspace{2em}}_{\text{total no. of boxes}}$

So many more "negatives" (boxes w. out cats) than "positives" (boxes with cats)

(3)

Precision & Recall

Spring 2013

Suppose the dataset has n_1 targets, and n_2 backgrounds
 - i.e. $\{(x_i, y_i) : i = 1 \text{ to } N\}$, $n_1 = \sum_{i=1}^N I(y_i = 1)$
 $n_2 = \sum_{i=1}^N I(y_i = -1)$ ← indicator function

We have a set of decision rules $d_T(\cdot)$
 which are parameterized by a threshold T .

i.e. each decision rule is of form

$$d_T(x) = 1, \text{ if } f(x) > T$$

$$= -1, \text{ if } f(x) < T$$

(example of $f(x)$ on next page)

Let $m_1^T =$ true +ves with $d_T =$ number of targets (rats) detected by d_T
 $m_2^T =$ false +ves with $d_T =$ number of backgrounds (non-rats) detected (incorrectly) by d_T

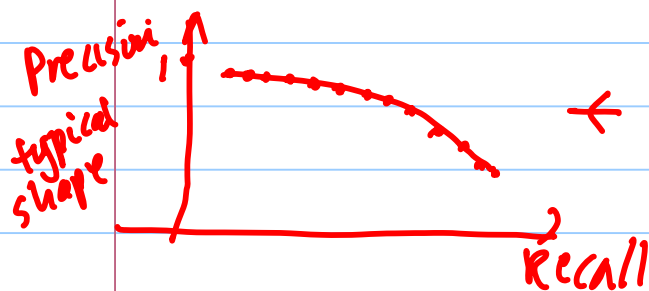
i.e. $m_1^T = \sum_{i=1}^N I(d_T(x_i) = 1) I(y_i = 1)$

$m_2^T = \sum_{i=1}^N I(d_T(x_i) = -1) I(y_i = -1)$

$I(\cdot)$
indicator function

Precision - at threshold $T = \frac{m_1^T}{m_1^T + m_2^T} = \frac{\text{No. True +ves}}{\text{No. of +ves}}$

Recall - at threshold $T = \frac{m_1^T}{n_1^T} = \text{Proportion of targets that are detected.}$
 (= No. of true +ves + No. of false +ves)



← Plot curve - each point corresponds to precision/recall at a value of T .
 Trade-off. High Precision - Low Recall
 Low Precision - High Recall

(4) ROC Curves

Note Title

10/3/2006

Bayes Decision Theory. $R(\alpha) = \sum_{x,y} L(\alpha(x), y) P(x, y)$

For binary $y \in \mathcal{Y}$ (i.e. $y \in \{\pm 1\}$) x, y

The Bayes Rule $\hat{\alpha} = \text{Arg min}_{\alpha} R(\alpha)$

reduces to thresholding the log-likelihood ratio. i.e. it is of form:

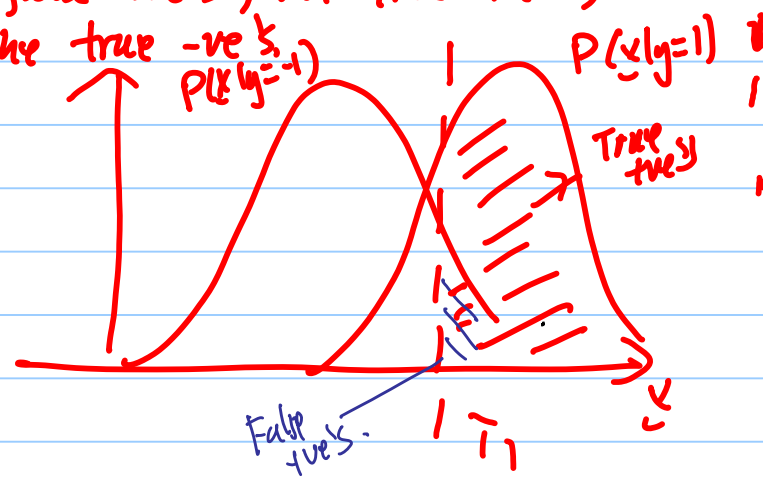
$$\hat{\alpha}_T(x) = 1, \text{ if } \log \frac{P(y=1|x)}{P(y=-1|x)} > T$$

$$\hat{\alpha}_T(x) = -1, \text{ otherwise}$$

The threshold T is a function of the prior $P(y)$ and the loss function $L(\alpha(x), y)$. Hence changing the prior, or the loss function, corresponds to changing T .

So $\log \frac{P(y=1|x)}{P(y=-1|x)}$ is an example of the function $f(x)$ (previous page)

Changing T will alter the false +ve's, the true +ve's, the false -ve's, the true -ve's.



(5)

Spring 2013

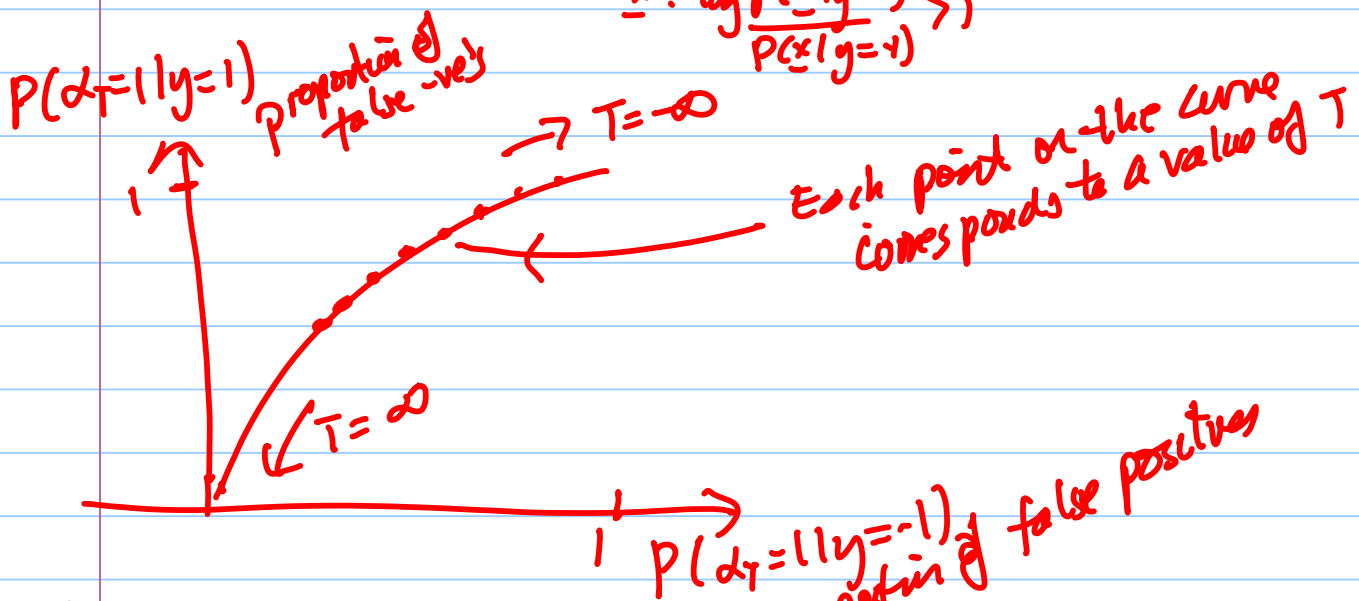
ROC curve. Plot $P(\alpha_T = 1 | y = 1)$
versus $P(\alpha_T = 1 | y = -1)$

$$P(\alpha_T = 1 | y = 1) = \sum_{\underline{x}} P(\alpha_T(\underline{x}) = 1 | \underline{x}) P(\underline{x} | y = 1)$$

$$= \sum_{\underline{x}: \log \frac{P(\underline{x} | y = 1)}{P(\underline{x} | y = -1)} > T} P(\underline{x} | y = 1)$$

$$P(\alpha_T = -1 | y = 1) = \sum_{\underline{x}} P(\alpha_T(\underline{x}) = -1 | \underline{x}) P(\underline{x} | y = -1)$$

$$= \sum_{\underline{x}: \log \frac{P(\underline{x} | y = 1)}{P(\underline{x} | y = -1)} > T} P(\underline{x} | y = -1)$$



Rule $\alpha(\underline{x}) = 1$ if $\log \frac{P(\underline{x} | y = 1)}{P(\underline{x} | y = -1)} > T$

So if $T = -\infty$, then all data is classified as positive
so $P(\alpha_T = 1 | y = -1) = P(\alpha_T = 1 | y = 1) = 1$

If $T = \infty$, all data is classified as negative $P(\alpha_T = 1 | y = -1) = P(\alpha_T = 1 | y = 1) = 0$
Bayes decision is given by a specific point T^* on the curve.

Spring 2013

(6)

The Curse of Dimensionality

The examples of Bayes Decision theory are misleading because they are given in low-dimensional spaces (1-dim, or 2-dim)

Many pattern classification tasks occur in high dimensional spaces. In these spaces our geometric intuitions are often wrong.

EG. Consider the volume of a sphere of radius $r=1$ in D dimension.

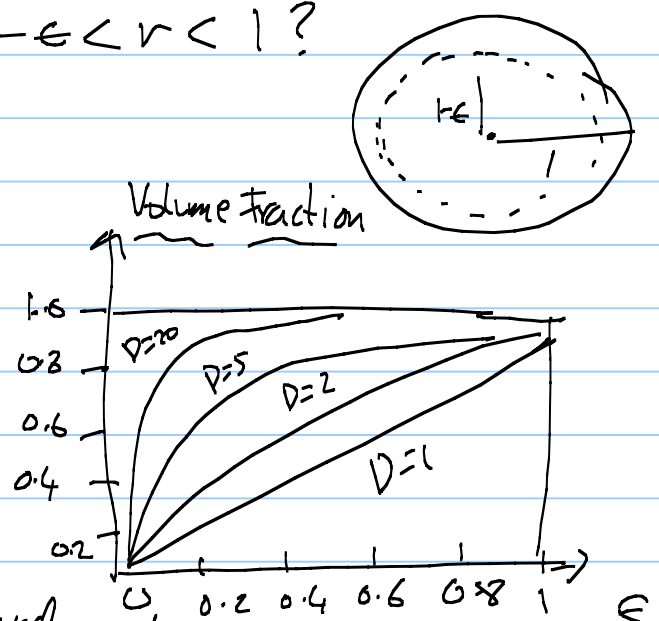
What fraction of its volume lies in the region between $1-\epsilon < r < 1$?

$$V_D(r) = k_D r^D$$

$$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = 1 - (1-\epsilon)^D$$

For large D , the volume fraction tends to 1 even for small ϵ .

Most of the volume is at the boundary!

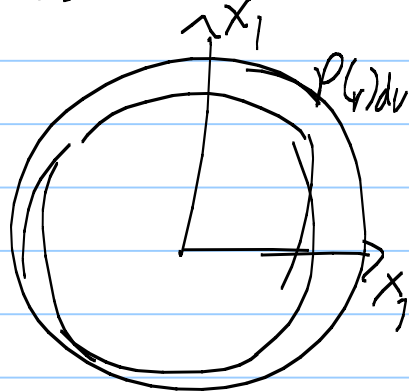


Spring 2013

(7)

e.g. Behaviour of a Gaussian distribution.

In 1-D, $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}$



In 2-D

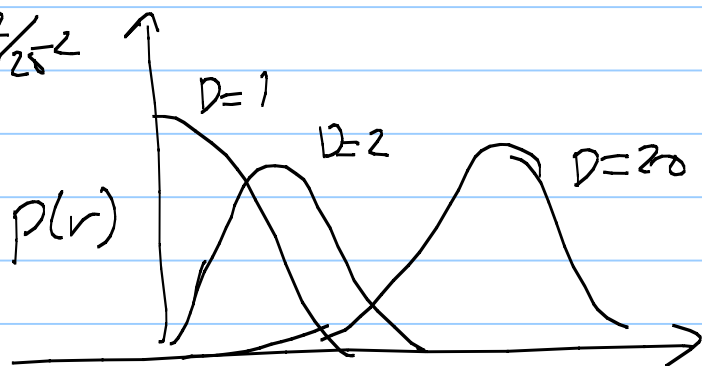
$$p(x_1, x_2) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x_1^2 + x_2^2)}{2\sigma^2}}$$

Let $r = \sqrt{x_1^2 + x_2^2}$.

Then $p(r) = \frac{r}{2\pi\sigma^2} e^{-r^2/2\sigma^2}$

In higher dimensions

$$p(r) = \frac{r^{D-1}}{K} e^{-r^2/2\sigma^2}$$



So in high dimensions most of the probability

mass of the Gaussian is

concentrated on a thin shell away from the center of the Gaussian.

Spring 2013

(8)

Learning probability distributions in high dimension can require a lot of data.

E.G. Gaussian Distribution in D dimension.

mean - μ D dimension.

covariance - Σ $D(D+1)/2$ dimension.

This is $O(D^2)$, not too bad.

But suppose we represent the data by a histogram with B bins per dimension.

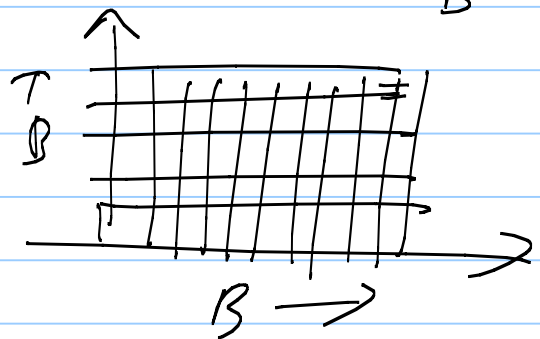
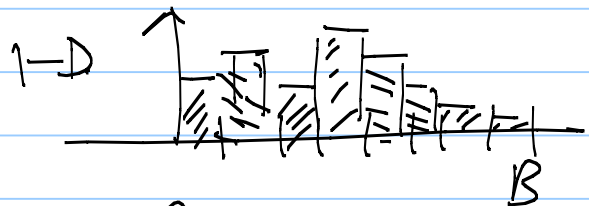
B bins in $D=1$

B^2 bins in $D=2$

B^D bins in D dimension.

Exponential growth!

Requires exponential amount of data to learn the distribution.



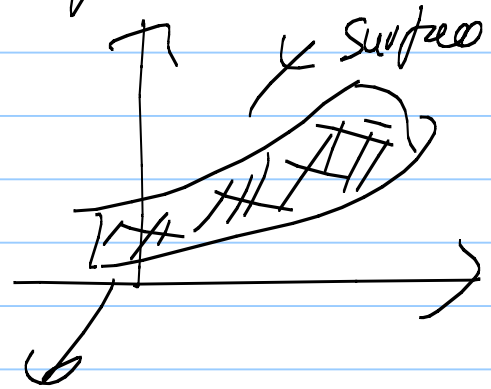
Spring 2013

(9)

How to deal with the curse of dimensionality?

In practice, data typically lies on some low-dimensional surface in the high dimensional space.

So the effective dimension of the data may be a lot smaller than the dimension of the space.



(*) Dimension Reduction Methods attempt to reduce the dimension by seeking this low dimensional surface. (Not always easy).

(*) Modeling, if we can guess distributions for the data (e.g. Gaussian) then dependence on the dimension is not too bad.

(*) Concentrate on the Decision Boundary ~ there may be enough data to learn the decision boundary even if we cannot learn the distributions.

(1b)

Bias and Variance

Spring 2013

This is a classical statistics perspective on generalization. First we need to introduce some statistics terminology.

Suppose we want to estimate a continuous quantity θ - e.g. the mean/variance of a Gaussian distribution (more about this in the next lecture), or the parameter of a regression line (see below) - then statisticians use an estimator.

The estimator is based on a set $X = \langle x_i; i=1 \dots N \rangle$ of examples - drawn from an unknown distribution $P(x)$.

$$\text{i.i.d. } P(X) = \prod_{i=1}^N P(x_i)$$

The task is to estimate a property θ by an estimator $\hat{\theta} = g(X)$. E.g. like a classification rate - but based on the set X , and θ is continuous.

For example: let $\theta = (\mu, \sigma)$ be the mean and variance of the data (data is one-dimensional in this example)

$$\text{then } X = \langle x_i; i=1 \dots N \rangle$$
$$\hat{\mu}(X) = \frac{1}{N} \sum_{i=1}^N x_i, \quad \hat{\sigma}^2(X) = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu}(X))^2$$

$$g(X) = (\hat{\mu}(X), \hat{\sigma}^2(X))$$

Note: that the estimator is a function of the set X , this will be important later

Spring 2013.

(11)

To evaluate the estimator, we want to measure how much it differs from the correct θ . It is attractive to use a quadratic errors (this helps the analysis)

$$(g(x) - \theta)^2 \quad \text{— but this depends on the data set } \mathcal{X}$$

So we need to get the expected error with respect to the set \mathcal{X} ,

$$r(g, \theta) = E_{\mathcal{X}} [g(x) - \theta]^2 = \int (g(x) - \theta)^2 P(x) dx$$

mean square error. distribution \mathcal{X} over set \mathcal{X} (i.e.).

$$b_{\theta}(g) = E_{\mathcal{X}} [g(x)] - \theta, \quad \text{bias of estimator}$$

If $b_{\theta}(g) = 0$ for all θ , then g is an unbiased estimator of θ

Eg. Consider $\hat{\mu}(x) = \frac{1}{N} \sum_{i=1}^N x_i$ estimator of the mean.

$$E_{\mathcal{X}} [\hat{\mu}(x)] = E_{\mathcal{X}} \left[\frac{1}{N} \sum_{i=1}^N x_i \right] = \frac{1}{N} \sum_{i=1}^N E_{x_i}(x_i) = \frac{1}{N} \sum_{x_i} x_i P(x_i) = \mu$$

Hence $\hat{\mu}(\cdot)$ is an unbiased estimator of μ . $\hat{\mu}(x)$ will depend on \mathcal{X} , but on average it gives you the right answer. mean of the distribution that generates the data

We can compute the variance of the estimator — i.e. how much it varies depending on \mathcal{X} .

(12)

Process $P(x)$ Spring 2013

$\mathcal{X}_1 = \{x_1, \dots, x_N\}$ $\mathcal{X}_2 = \{x_{N+1}, \dots, x_{2N}\}$ $\mathcal{X}_3 = \{x_{2N+1}, \dots, x_{3N}\}$
 possible datasets

$g(\mathcal{X}_1), g(\mathcal{X}_2), g(\mathcal{X}_3)$, these will vary.

Calculate the mean \rightarrow e.g. $E_{\mathcal{X}}\{g(\mathcal{X})\}$

Calculate the variance $\rightarrow \text{Var}(g(\mathcal{X})) = E_{\mathcal{X}}\{(g(\mathcal{X}) - E_{\mathcal{X}}\{g(\mathcal{X})\})^2\}$.

For the estimator $\hat{\mu}(\mathcal{X})$ of the mean,
 we know that $E_{\mathcal{X}}\{\hat{\mu}(\mathcal{X})\} = \mu$ unbiased (previous page)

$$\begin{aligned} \text{Var}_{\mathcal{X}}(\hat{\mu}) &= \text{Var}_{\mathcal{X}}\left(\frac{\sum_{i=1}^N x_i}{N}\right) = \frac{1}{N^2} \sum_{i=1}^N \text{Var}_{x_i}(x_i) \\ &= \frac{N}{N^2} \sigma^2 = \frac{\sigma^2}{N} \end{aligned}$$

$\underbrace{\sum_{x_i} P(x_i)}_{x_i} (x_i - \mu)^2$

Hence, the variance of the estimator tends to zero as $N \rightarrow \infty$ will fall off rate $1/N$

Note: this $1/N$ fall-off rate is true for any linear estimator - i.e. $g(\mathcal{X})$ is a linear function of the elements x_1, \dots, x_N in set \mathcal{X} .

Next, consider the bias of $\hat{\sigma}^2(\mathcal{X}) = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu}_i)^2$

By similar analysis (to analysis of $\hat{\mu}(\mathcal{X})$) we find that
 $E_{\mathcal{X}}[\hat{\sigma}^2(\mathcal{X})] = \frac{1}{N} \left(\sum_{i=1}^N E_{x_i}(x_i^2) - N\hat{\mu}^2 \right) = \frac{N-1}{N} E_{\mathcal{X}}\{(x-\mu)^2\} = \frac{N-1}{N} \sigma^2$
 (uses $\sum_{i=1}^N (x_i - \hat{\mu})^2 = \sum_{i=1}^N x_i^2 - N\hat{\mu}^2$). biased, but asymptotically unbiased.

(13)

Bias Variance Dilemma:

Spring 2013

Note Title

Adjusting Model Complexity

3/29/2008

Dataset

$$\mathcal{X} = \langle (x_i, y_i) : i = 1 \text{ to } n \rangle,$$

sampled from $P(x, y) = p(y|x)P(x)$

y is continuous valued.

Let $g(x)$ be an estimator of y

Claim 1: $\langle (y - g(x))^2 \rangle_{P(y|x)} = \underbrace{\langle y - \langle y \rangle_{P(y|x)} \rangle_{P(y|x)}^2}_{\text{Expected error w.r.t. } P(y|x)} + \underbrace{\langle \langle y \rangle_{P(y|x)} - g(x) \rangle_{P(y|x)}^2}_{\text{Squared error}}$

Expected error w.r.t. $P(y|x)$

Variance of the process

Squared error

→ Nothing we can do about this - indep of our estimator $g(\cdot)$

$$\begin{aligned} \text{Here } & \langle f(y, x) \rangle_{P(y|x)} \\ &= \sum_y f(y, x) P(y|x) \end{aligned}$$

Proof: write $(y - g(x))^2 = (y - \langle y \rangle_{P(y|x)} + \langle y \rangle_{P(y|x)} - g(x))^2$

$$= (y - \langle y \rangle_{P(y|x)})^2 + (\langle y \rangle_{P(y|x)} - g(x))^2 + 2(y - \langle y \rangle_{P(y|x)}) (\langle y \rangle_{P(y|x)} - g(x))$$

Take expectation with respect to $P(y|x)$

Then $\langle (y - g(x))^2 \rangle_{P(y|x)} = \langle (y - \langle y \rangle_{P(y|x)})^2 \rangle_{P(y|x)} + \langle \langle y \rangle_{P(y|x)} - g(x) \rangle_{P(y|x)}^2$

because $\langle \langle y \rangle_{P(y|x)} - g(x) \rangle_{P(y|x)}^2$ is indep of y and $\langle (y - \langle y \rangle_{P(y|x)}) \rangle_{P(y|x)} = 0$

Claim 1 says we can decompose the expected error (w.r.t. $P(y|x)$) into part we have no control over $\langle (y - \langle y \rangle_{P(y|x)})^2 \rangle_{P(y|x)}$ and part which depends on $g(\cdot)$ and the data \mathcal{X} .

$$\langle \langle y \rangle_{P(y|x)} - g(x) \rangle_{P(y|x)}^2$$

(14) Next we study the expectation Spring 2013

of $(\langle y \rangle_{P(y|x)} - g(x))^2$ with respect to $P(x)$

i.e. how it depends on the particular sample $X = (x_1, \dots, x_n)$ from $P(x)$.

Claim II $\langle (\langle y \rangle_{P(y|x)} - g(x))^2 \rangle_{P(x)} = \underbrace{(\langle y \rangle_{P(y|x)} - \langle g(x) \rangle_{P(x)})^2}_{\text{Bias}} + \underbrace{\langle (g(x) - \langle g(x) \rangle_{P(x)})^2 \rangle_{P(x)}}_{\text{Variance}}$

The Bias-Variance Result.

The first term depends on the bias - the difference between the best estimate $\langle y \rangle_{P(y|x)}$ (if we knew the distribution) and the expectation of our estimator $\langle g(x) \rangle_{P(x)}$

The second term is the variance of the estimator $g(x)$ - i.e. how much it depends on the sample set X (more labels)

Proof: Write $(\langle y \rangle_{P(y|x)} - g(x))^2$

$$= (\langle y \rangle_{P(y|x)} - \langle g(x) \rangle_{P(x)} + \langle g(x) \rangle_{P(x)} - g(x))^2$$

$$= (\langle y \rangle_{P(y|x)} - \langle g(x) \rangle_{P(x)})^2 + (\langle g(x) \rangle_{P(x)} - g(x))^2$$

$$+ 2(\langle y \rangle_{P(y|x)} - \langle g(x) \rangle_{P(x)})(\langle g(x) \rangle_{P(x)} - g(x))$$

Take expectation wrt. $P(x)$

$$\langle (\langle y \rangle_{P(y|x)} - g(x))^2 \rangle_{P(x)} = (\langle y \rangle_{P(y|x)} - \langle g(x) \rangle_{P(x)})^2 + \langle (g(x) - \langle g(x) \rangle_{P(x)})^2 \rangle_{P(x)}$$

Because $\langle y \rangle_{P(y|x)} - \langle g(x) \rangle_{P(x)}$ is independent of x

$$\text{and } \langle g(x) - \langle g(x) \rangle_{P(x)} \rangle_{P(x)} = 0$$

(15)

What does this mean?

Spring 2013.

Distribution $P(x)$

Dataset $X_1 = \{x_1, \dots, x_N\}$, $X_2 = \{x_{N+1}, \dots, x_{2N}\}$
 $\dots \dots X_m = \{x_{(m-1)N+1}, \dots, x_{mN}\}$

For each data we get an estimate of y

$g(x_1), g(x_2) \dots g(x_m)$

The mean estimate is $\bar{g} = \frac{1}{m} \sum_{i=1}^m g(x_i)$

the variance is $\text{Var}(\bar{g}) = \frac{1}{2} \sum_{i=1}^m (g(x_i) - \bar{g})^2$

To get good generalization we want the variance to be small, so that it isn't sensitive to the data we have trained the classifier on.

Ideally we want to have a classifier $g(\cdot)$ which has small bias and variance.

In practice, there is often a trade-off between bias and variance.

A complex classifier can give a good fit to the data (compared to a simple classifier)

but can have high variance because it over-fits the data. So it gives different results on different datasets.

116)

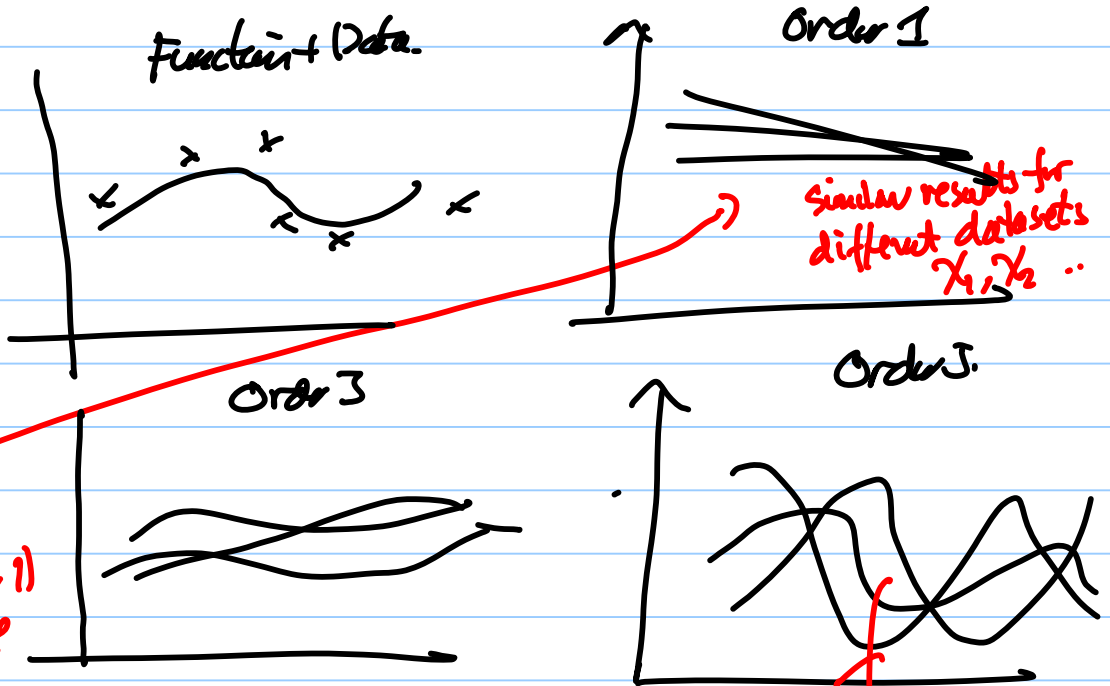
Spring 2017

Bias/Variance Dilemma

The data is generated by
 $f(x) = 2\sin(1.5x)$
 $\epsilon \sim N(0, 1)$

Fit an order 1 polynomial (straight line) to the data.

Fit order 3 poly. Better fit (than order 1) but more variance.



A more complex models gives better fit to the data (i.e. to underlying model) → reduces bias.

But small changes in dataset lead to big variance in fitted model → increases variance.

Low orders — risk of underfitting
High orders — risk of overfitting.

(fit the noise, not the function)

To get a small error — we should have the proper inductive bias and have large enough dataset so that variability is constrained by data.
Note: take many high-variance models, use average (labeled)

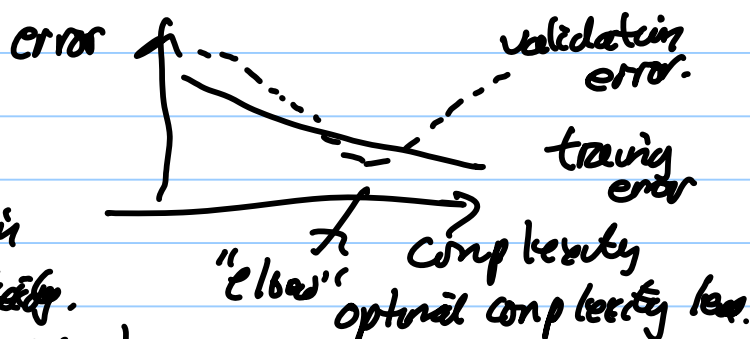
(17)

What to do? How to test generalization? Spring 2013

Cross-validation → divide dataset into two parts as training & validation set.

Train models of different complexity and test their error on the validation set.

As model complexity increases, error on training set decreases. But error on validation set decreases then increases.



regularization

augmented error function

$$\hat{E}' = \text{error on data} + \lambda \cdot \text{model complexity.}$$

(λ optimized using cross-validation)

structural risk minimization (Vapnik)

(λ dim)

— also penalizes model complexity.

Minimum Description Length (Rissanen)

penalize complexity by cost of encoding model.

Bayesian Model Selection. if some prior knowledge

$$P(\text{model} | \text{data}) = \frac{P(\text{data} | \text{model}) P(\text{model})}{P(\text{data})}$$

(gives higher prob to simpler models)