1 Lecture 3 Spring 2014

Note Title 4/10/2013

Three Topics:

(I) Receiver Operating Characteristic (ROC) curves
Precision and Recall Curves.
What if we do not select a 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

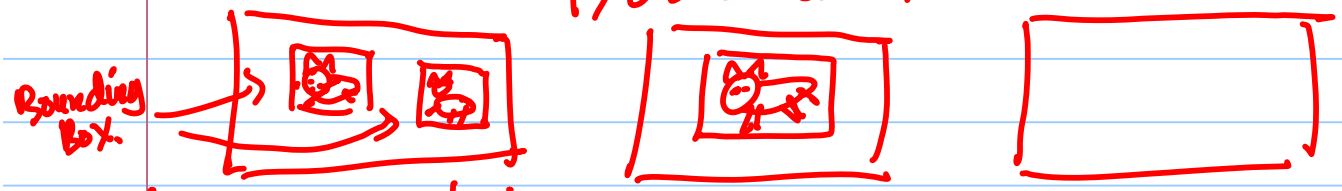
III The Bias-Variance Dilemma.
* A classic statistical perspective about generalization.

(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 seperate them into groups for further testing.

Example: Detecting Cats in the Pascal Challenge (Computer Vision)

Data: 20,000 images
1,000 cats.



Each image can contain 0 cats, 1 cat, or ≥2 cats.

The positions 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 $20{,}000 \times 1{,}000 = 20{,}000{,}000$
  (no. images; no. boxes in each image; total no. of boxes.)

So many more "negatives" (boxes without cats) than "positives" (boxes with cats)

(3) # Precision & Recall

Note: $I(y=1) = 1$, if $y=1$; Indicator Function $= 0$, if $y \neq 1$

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

$n_2 = \sum_{i=1}^{N} I(y_i = -1)$.

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

i.e. each decision rule is of form

$\alpha_T(x) = 1$, if $f(x) > T$

$= -1$, if $f(x) < T$

(if rule $\alpha_T(.)$ is used) (example of $f(x)$ on next page)

Let m_1^T = true positives = number of targets (cats) detected by α_T

m_2^T = false positives = number of backgrounds (non-cats) detected (incorrectly) by α_T

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

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

$I(..)$ indicator function

Precision – at threshold T $= \dfrac{m_1^T}{m_1^T + m_2^T}$ = $\dfrac{\text{No. True +ve's}}{\text{No. of +ve's}}$ ($=$ No. of (True +ves) + No. of (False +ves))

Recall – at threshold T $= \dfrac{m_1^T}{n_1^T}$ = Proportion of Targets that are detected.

← 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

Spring 2013.

Note Title 10/3/2006

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

For binary $y \in \{\pm 1\}$

The Bayes Rule $\hat{\alpha} = \arg\min_{\alpha} R(\alpha)$

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

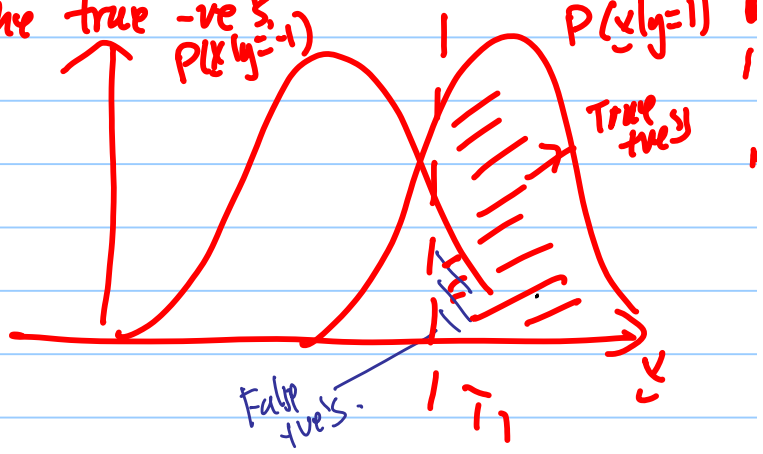
$$\hat{\alpha}_T(x) = 1 \text{ , if } \log \frac{P(y=1|\underline{x})}{P(y=-1|\underline{x})} > T$$

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

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

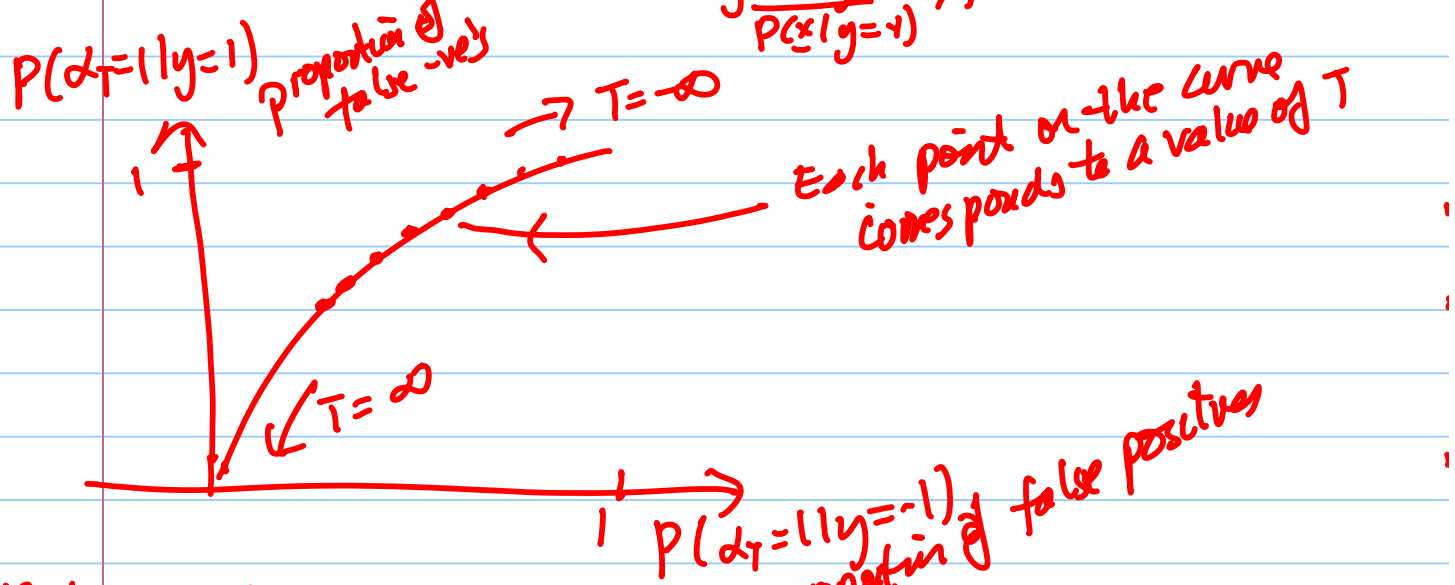
ROC curve. Plot $P(\alpha_T = 1 | y=1)$ versus $P(\alpha_T = 1 | y=-1)$

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

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

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

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



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.

(6)

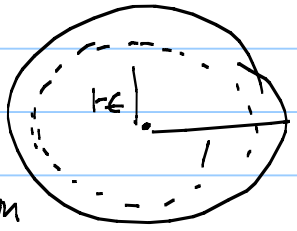
The Curse of Dimensionality

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.

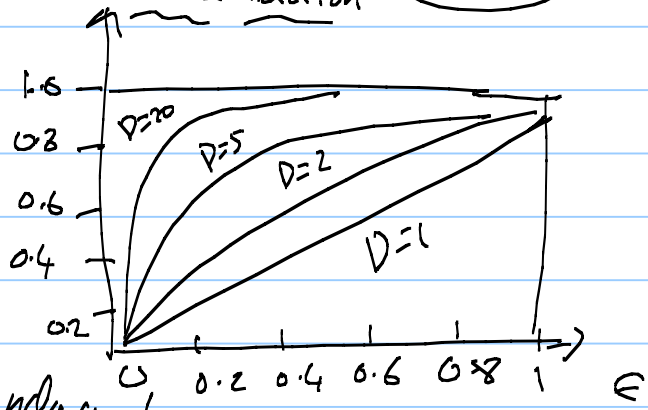
Eg. Consider the volume of a sphere of radius $r=1$ in D dimension.

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



$V_D(r) = K_D r^D$

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

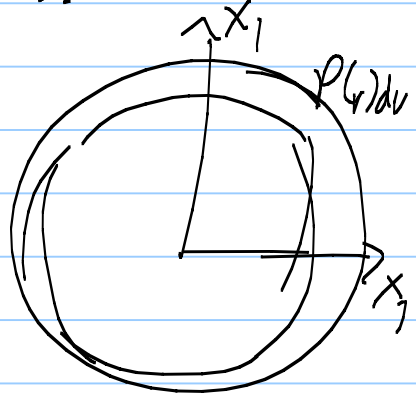


For large D, the volume fraction tends to 1 even for small ϵ.

Most of the volume is at the boundary!

(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

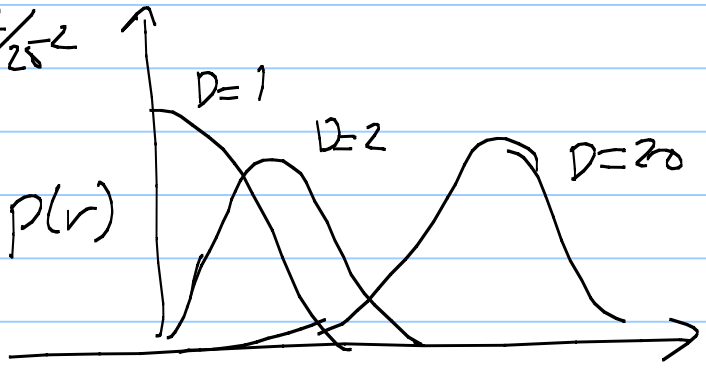
$$p(x_1, x_2) = \frac{1}{2\pi\sigma^2} e^{-(x_1^2+x_2^2)/2\sigma^2}.$$

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.

(8)

Learning probabilities distributions in high dimension can require a lot of data

E.G. Gaussian Distribution in D dimension.

mean – $\underline{\mu}$ D dimensions.

Covariance – $\underline{\underline{\Sigma}}$ $\frac{D(D+1)}{2}$ Dimension.

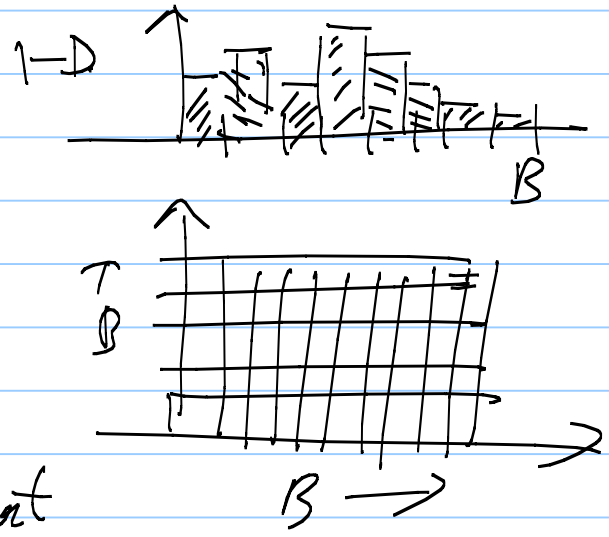
This is $O(D^2)$, not too bad.

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.



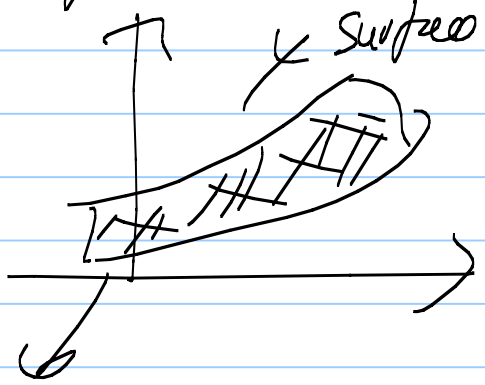
Exponential growth!

Requires exponential amount of data to learn the distribution.

19)

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 the 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.

(16) Bias and Variance

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 parameters of a regression line (see below) – then statisticians use an estimator.

The estimator is based on a set $X_N = \{\underline{x}_i : i = 1..N\}$ of examples – drawn from an unknown distribution $P(\underline{x})$.

i.i.d. $P(X_N) = \prod_{i=1}^{N} P(\underline{x}_i)$

The task is to estimate a property θ by an estimator $\hat{\theta} = g(X_N)$. Eg. like a classification rule – but based on the set X_N 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)

then $X_N = \{x_i : i = 1..N\}$

$$\hat{\mu}(X_N) = \frac{1}{N}\sum_{i=1}^{N} x_i \,, \quad \hat{\sigma}^2(X_N) = \frac{1}{N}\sum_{i=1}^{N} (x_i - \hat{\mu}(X_N))^2$$

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

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

(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)

$(g(\chi_N) - \theta)^2$ – but this depends on the data set χ_N

So we need to get the expected error with respect to the set χ_N,

$$r(g,\theta) := E_\chi[g(\chi) - \theta]^2 = \int (g(\chi) - \theta)^2 P(\chi) d\chi$$

mean square error. distribution over set $\chi = \{x_i\}$.

$b_\theta(g) = E_\chi[g(\chi)] - \theta$, bias of estimator

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

Eg. consider $\hat{\mu}(\chi) = \frac{1}{N}\sum_{i=1}^{N} x_i$, estimate of the mean

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

$$= \sum_{x_i} x_i P(x_i) = \mu$$

mean of the distribution that generated the data.

$$= \frac{1}{N} N\mu = \mu$$

Hence $\hat{\mu}(\cdot)$ is an unbiased estimator of μ.

($\hat{\mu}(\chi)$ will depend on χ, but on average it gives you the right answer).

We can compute the variance of the estimator – i.e. how much it varies depending on χ.

Generating Process $P(X)$

$P(X)$ can generate many possible datasets $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3 \ldots$

$\mathcal{X}_1 = \{x_1 \ldots x_N\}$ $\quad \mathcal{X}_2 = \{x_{N+1}, \ldots, x_{2N}\}$ $\quad \mathcal{X}_3 = \{x_{2N+1}, \ldots, x_{3N}\}$

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

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

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

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

$$\mathrm{Var}_{\mathcal{X}}(\hat{\mu}) = \mathrm{Var}_{\mathcal{X}}\left(\frac{\sum_{i=1}^{N} x_i}{N}\right) = \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{Var}_{x_i}(x_i) = \sum_{x_i} P(x_i)(x_i - \mu)^2$$

$$= \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$$

Hence, the variance of the estimator tends to zero as $N \to \infty$ with 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 $x_1 \ldots x_N$ 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

$E_{\mathcal{X}}[\hat{\sigma}^2(\mathcal{X})] = \frac{1}{N}\left\{\sum_{i=1}^{N} E_{x_i}(x_i^2) - N\hat{\mu}^2\right\} = \frac{N-1}{N} E_x((x-\mu)^2) = \frac{N-1}{N}\sigma^2$

(uses $\sum_{i=1}^{N}(x_i - \hat{\mu})^2 = \sum_{i=1}^{N} x_i^2 - \hat{\mu}^2$).

biased, but asymptotically unbiased. as $N \to \infty$

(13) Bias Variance Dilemma:

Note Title 3/29/2008

Dataset $\chi = \{(x_i, y_i) : i = 1 \text{ to } N\}$
sampled from $P(x,y) = p(y|x)p(x)$
y is continuous valued.

Let $g(\chi)$ be an estimator of y

Claim I: $\langle (y - g(\chi))^2 \rangle_{P(y|\chi)} = (y - \langle y \rangle_{P(y|\chi)})^2 + (\langle y \rangle_{P(y|\chi)} - g(\chi))^2$

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

Variance of the process → Nothing we can do about this - independent of our estimator $g(\cdot)$

squared error

Here $\langle f(y,\chi) \rangle_{P(y|\chi)} = \sum_y f(y,\chi) P(y|\chi)$

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

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

Take expectation with respect to $P(y|\chi)$

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

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

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

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

(14) Next we study the expectation of $(\langle y\rangle_{p(y|x)} - g(x))^2$ with respect to $P(X)$ i.e. how it depends on the particular sample $X = (x_1, \ldots x_N)$ from $P(x)$.

Claim II $\left\langle (\langle y\rangle_{p(y|x)} - g(x))^2 \right\rangle_{P(X)} = (\langle y\rangle_{p(y|x)} - \langle g(x)\rangle_{P(x)})^2 + \left\langle (g(x) - \langle g(x)\rangle_{P(x)})^2 \right\rangle_{P(x)}$

The Bias-Variance Result.

bias

variance

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 (noise level)

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 $\langle y\rangle_{p(y|x)} - g(x)\rangle_{P(x)}$ is independent of x and $\langle g(x) - \langle g(x)\rangle_{P(x)}\rangle_{P(x)} = 0$

(15)

What does this mean?

Distribution $P(x)$

Dataset $\mathcal{X}_1 = \langle x_1 \ldots x_N \rangle$, $\mathcal{X}_2 = \langle x_{N+1}, \ldots x_{2N} \rangle$

$\ldots \quad \mathcal{X}_m = \langle x_{(m-1)N+1}, \ldots x_{mN} \rangle$

For each data we get an estimate of y

$g(\mathcal{X}_1), g(\mathcal{X}_2) \ldots g(\mathcal{X}_m)$

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

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

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 $g(.)$ 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.

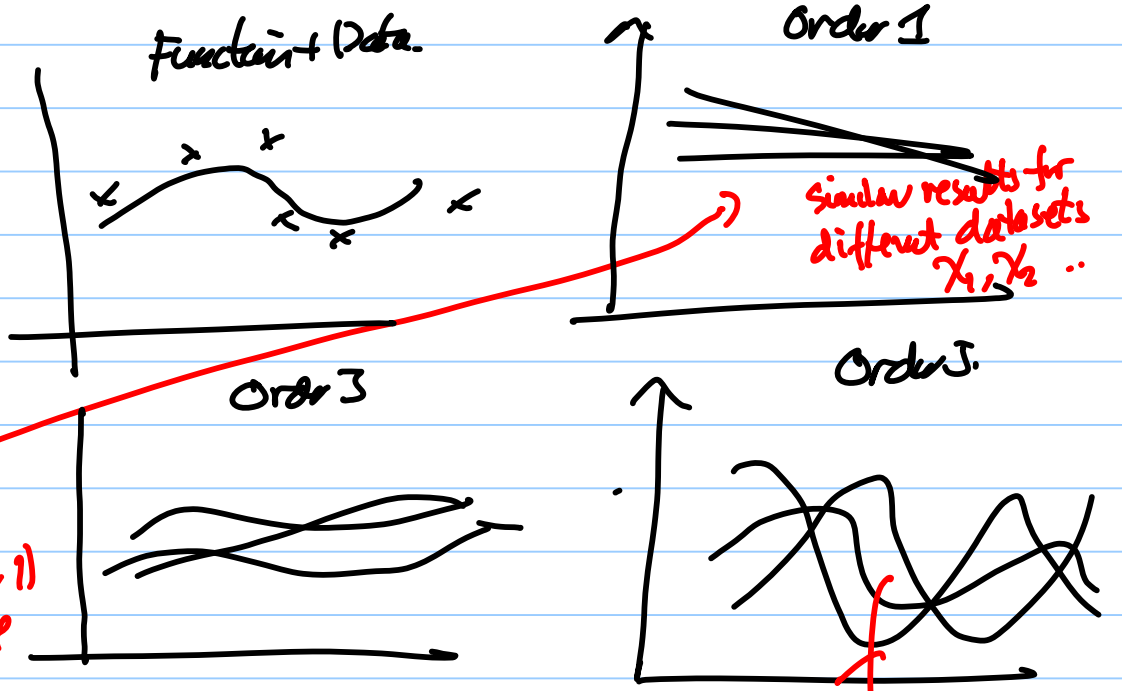
(16)

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 bigger variations 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 (later)

(17)

What to do? How to test generalization?

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

E' = error on data + λ · model complexity.

(λ optimized using cross-validation)

Structural risk minimization (Vapnik) (VC 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)