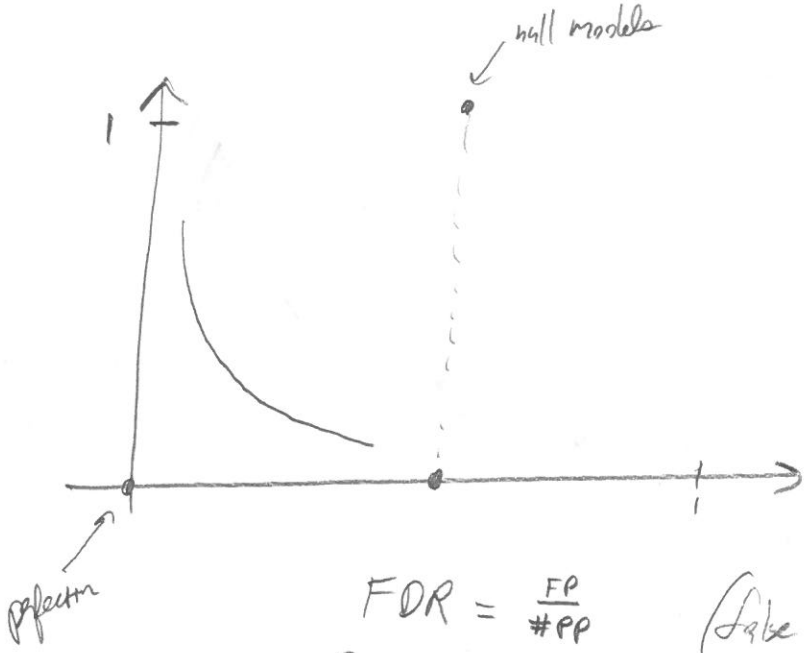
This is the most popular illustration of all model options, but not the only one. Also there's detection error tradeoff (DET, 1997)




$$FDR = \frac{FP}{\#PP} \quad \text{(false alarm)}$$

$(\neq FPR)$

Review

$\hat{y} = \mathbb{1}_{p \geq p_{th}}$ ← param. which determines tradeoff of FP & FN.

Look at all choices of $p_{th}$ in a "performance curve"

ROC


AUC: evaluates the all-area of classification models from the prob. classifier ⇒ evaluates prob. classifier itself

Under the null model

$$FDR = \frac{p_{th}(1-p_Y)n}{p_{th}n} = 1 - p_Y \quad \text{(always)}$$

As FP ↑ ⇒ FN ↓

$$\text{Miss Rate} = \frac{(1-p_{th})p_Y n}{p_Y n} = 1 - p_{th}$$

if $p_{th} = 0 \Rightarrow$ all $\hat{y}_i = 1$ Never misses anything!

if $p_{th} = 1 \Rightarrow$ all $\hat{y}_i = 0$ misses everything FDR undefined

---

How to select a classifier — One pt on the ROC or DET plot! You need to define your costs / rewards

$$\text{Avg Cost} = \frac{1}{n}\left(C_{FP}FP + C_{FN}FN\right)$$

$C_{FP}$ = cost of FP (as a #), $C_{FN}$ = cost of FN (as a #)

$$\text{Avg Reward} = \frac{1}{n}\left(-C_{FP}FP + r_{TP}TP + -C_{FN}FN + r_{TN}TN\right)$$

For example: a mailing advertisement

Imagine you are trying to send people $10 open a credit card and the mailing costs $5.

TP: you mailed a person who opens the new CC

$r_{TP}$ = $1,000 − $5 (mailing)

TN: you don't mail ..... doesn't open a new CC

$r_{TN}$ = $0

FP: you mail but the person doesn't open the C.C.

$C_{FP}$: $5

FN: you don't mail someone who would open the C.C.

$C_{FN}$: $1,000

So you want to build a model where FN are minimized and you don't care about FP's as much. $P_{FN}$ = small!

This is why you always get C.C. mailings! Because they don't care that you're a FP!!

DEMO

# Bias-Variance Decomposition

Very theoretical...

Back to regression... recall

$$y = g + e = g + \underbrace{(f-g)}_{\text{error due to missp. \& est error}} + \underbrace{\delta}_{\text{error due to ignorance AKA "irreducible error"}}$$

$$\Rightarrow \underbrace{y-g}_{\text{measure of how good the model } g \text{ is}} = f-g+\delta \qquad (y-g = e)$$

$$e^2 = (f-g+\delta)^2$$

What if I wanted the "mean squared error (MSE)". This would mean the expectation of the sqd. residuals

In order to take an expectation, you need r.v.'s! Let's try this:

$$Y = f(\vec{x}) + \Delta$$

($Y$: r.v. for response; $\Delta$: r.v. for irr. error)

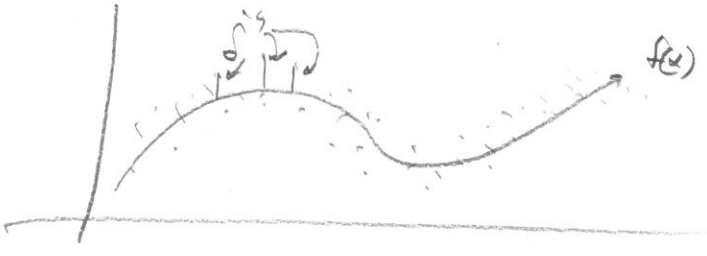
Does it really pass. How about... since $Y$ is a function of $\vec{x}$.

$$Y|\vec{X}=\vec{x} = f(\vec{x}) + \Delta|\vec{X}=\vec{x}$$

Economists call this the "conditional expectation function" (CEF)

Let's assume (I) $E[Y|\vec{X}=\vec{x}] = f(\vec{x})$ $\Rightarrow$ $E[\Delta|\vec{X}=\vec{x}] = 0$

(avg over irr. err)




So... for any $\vec{x}$, if many realizations of $Y$ $\Rightarrow$ the avg of those $y$'s is close to $f(\vec{x})$.

Variance does not depend on $\vec{x}$ (homoskedastic assumption)

(II) $Var[\Delta \mid \vec{X}=\vec{x}] = Var[\Delta] = \sigma^2 \Rightarrow E[\Delta^2] = 0$

---

Back to MSE, we care about this for a new obs. $\vec{x}^*$. This is also called "generalization error."

$$MSE(\vec{x}^*) := E\left[(Y^* - g(\vec{x}^*))^2 \mid \vec{X}=\vec{x}^*\right] \geq \sigma^2$$

can't do better than irreducible error

What if we know $f$?

$$MSE(\vec{x}^*) := E\left[(Y^* - f(\vec{x}^*))^2 \mid \vec{X}=\vec{x}^*\right] = E\left[\Delta^{*2} \mid \vec{X}=\vec{x}^*\right] = E[\Delta^2] = \sigma^2$$

What is this expectation over? $\Delta^*$. An integral over $\mathbb{R}$ with $P(\Delta)$ inside

If $f$ unknown...

$$MSE(\vec{x}^*) = E\left[Y^{*2} - 2Y^* g(\vec{x}^*) + g(\vec{x}^*)^2 \mid \vec{X}=\vec{x}^*\right]$$

$$= \underline{E\left[Y^{*2} \mid \vec{X}=\vec{x}^*\right]} - 2\,\underline{E\left[Y^* g(\vec{x}^*) \mid \vec{X}=\vec{x}^*\right]} + \underline{E\left[g(\vec{x}^*)^2 \mid \vec{X}=\vec{x}^*\right]}$$

$$= E\left[(f(\vec{x}) + \Delta)^2 \mid \vec{X}=\vec{x}^*\right] \qquad E[Y^*]\, g(\vec{x}^*) \qquad g(\vec{x})^2$$

$$= E\left[(f(\vec{x}) + \Delta)^2\right] \qquad = f(\vec{x}^*)\, g(\vec{x}^*)$$

$$= E\left[f(\vec{x})^2 + 2f(\vec{x})\Delta + \Delta^2\right]$$

$$= f(\vec{x})^2 + \sigma^2 - 2f(\vec{x}^*)g(\vec{x}^*) + g(\vec{x})^2$$

$$= \left(f(\vec{x}^*) - g(\vec{x}^*)\right)^2 + \sigma^2$$

expand sqd error as additive

$$e = (f-g) + f$$

Not so interesting... How about we take expectation over all $\Delta_1, \Delta_2, \ldots, \Delta_n$ and $\Delta^*$? Expectation over all $\mathbb{D}$!

$$MSE(\vec{x}^u) = E_{\Delta_1,\dots,\Delta_n,\Delta^u}\left[ Y^{u2} \mid \vec{X} = x^u \right] - 2\, E_{\Delta_1,\dots,\Delta_n,\Delta^u}\left[ Y^u g(\vec{x}^u) \mid \vec{X} = x^u \right] + E_{\Delta_1,\dots,\Delta_n}\left[ g(\vec{x}^u)^2 \mid \vec{x}^u \right]$$

($Y^{u2}$: only varies with $\Delta^u$)

$$A(\mathcal{D}, \mathcal{U}) = R(X, \vec{y}, \mathcal{U})$$

which is a function of $\Delta$'s! but not a function of $x$

$$= E_{\Delta^u}\left[ Y^{u2} \mid \vec{X} = \vec{x}^u \right] - 2\, E_{\Delta_1,\dots,\Delta_n,\Delta^u}\left[ Y^u g(\vec{x}^u) \mid \vec{X} = \vec{x}^u \right] + E_{\Delta_1,\dots,\Delta_n}\left[ g(\vec{x}^u)^2 \right]$$

($g$: independent)

$$= \sigma^2 + f(\vec{x}^u)^2 - 2\, \underbrace{E_{\Delta^u}[Y^u]}_{f(\vec{x}^u)}\, E_{\Delta_1,\dots,\Delta_n}\left[ g(\vec{x}^u) \right] + E_{\Delta_1,\dots,\Delta_n}\left[ g(\vec{x}^u)^2 \right]$$

$$= \sigma^2 + Var\left[ g(\vec{x}^u) \right] + f(\vec{x}^u)^2 - 2 f(\vec{x}^u) E\left[ g(\vec{x}^u) \right] + E\left[ g(\vec{x}^u) \right]^2$$

($Var[g(\vec{x}^u)] + E[g(\vec{x}^u)]^2$)

$$= \sigma^2 + Var\left[ g(\vec{x}^u) \right] + \left( E\left[ g(\vec{x}^u) \right] - f(\vec{x}^u) \right)^2$$

$$= \sigma^2 + \underbrace{Var\left[ g(\vec{x}^u) \right]}_{E\left[ g(\vec{x}) - E(\vec{x}) \right]} + \underbrace{E\left[ g(\vec{x}^u) - f(\vec{x}^u) \right]^2}_{Bias\left[ g(\vec{x}) \right]^2}$$

How much $g$ varies around its mean (system $E(\vec{x})$) (instability in learning)

How far away the avg $g$ is from $f$.

Bias-Var tradeoff

One more small pt. This was all for one data pt $\vec{x}^u$. We can average over all $\vec{x}_1, \dots, \vec{x}_n \; \vec{x}^u \in \mathcal{X}$. Assume distr. $P(\vec{x})$.

$$MSE = E_{\mathcal{X}}\left[ \sigma^2 + Var\left[ g(\vec{x}^u) \right] + Bias\left[ g(\vec{x}) \right]^2 \right]$$

(help of $\mathcal{X}$)

Since there one $\mathcal{D}$ and one $g$... we don't get to see this except during simulation

$$= \sigma^2 + E_{\mathcal{X}}\left[ Var\left[ g(\vec{x}^u) \right] \right] + E_{\mathcal{X}}\left[ Bias\, g(\vec{x}^u)^2 \right]$$

gen. error.

expected Variance

expected $Bias^2$

demo