Lecture 7

"Plug-in" approximation

Consider predictive prob. distribution:

$$p(\vec{y}|\vec{x},\mathcal{D}) = \int d\vec{\theta}\, p(\vec{y}|\vec{x},\vec{\theta})\, p(\vec{\theta}|\mathcal{D}).$$

$\int d\vec{\theta}$ hard to compute in general

If we replace $p(\vec{\theta}|\mathcal{D})$ by $\delta(\vec{\theta}-\vec{\theta}^{MAP})$, [or even $\delta(\vec{\theta}-\vec{\theta}^{*})$] $\leftarrow$ MLE

we obtain:

$$p(\vec{y}|\vec{x},\mathcal{D}) \simeq p(\vec{y}|\vec{x},\vec{\theta}^{MAP}).$$

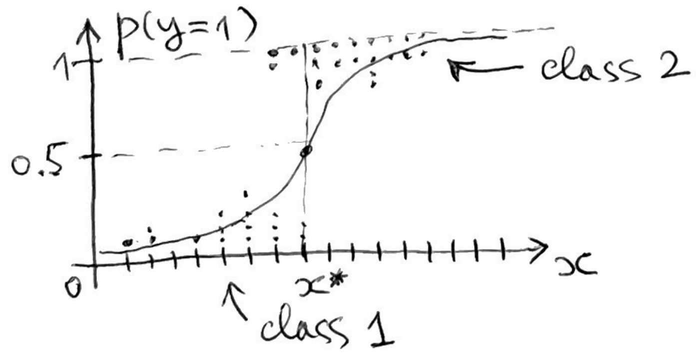
<u>Ex.</u> binary classification, 1D inputs
$y \in \{0,1\}$

$$p(y|x;\vec{\theta}) = \mathrm{Ber}(y|\sigma(wx+b)).$$
$$\vec{\theta} = \{w, b\}$$

In other words,

$$p(y=1|x;\vec{\theta}) = \sigma(b+wx) = \frac{1}{1+e^{-(b+wx)}}.$$

If we fit $\vec{\theta}$ by ML on some data $\mathcal{D} = \{x_n, y_n\}_{n=1}^{N}$, we obtain:



Decision boundary (DB): value of x^* s.t.

$$p(y=1|x^*;\vec{\theta}^*) = p(y=0|x^*;\vec{\theta}^*) = 0.5 .$$

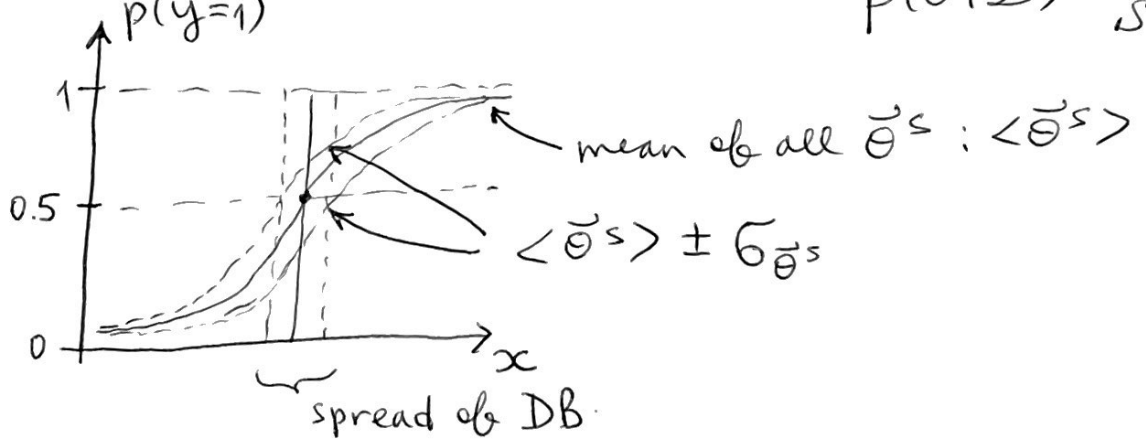
Here, $\sigma(b^* + w^* x^*) = 0.5 \Rightarrow b^* + w^* x^* = 0$, or

$$x^* = -\frac{b^*}{w^*} .$$

A better approach would be to integrate over $p(\vec{\theta}|\mathcal{D})$. In practice, this can be accomplished using Monte-Carlo (MC) approximation:

1. Draw S samples from the posterior:
$$\vec{\theta}^s \sim p(\vec{\theta}|\mathcal{D}) \qquad s = 1, \dots, S$$

2. Approx. pred. prob. as
$$p(y=1|x,\mathcal{D}) \simeq \frac{1}{S}\sum_{s=1}^{S} p(y=1|x,\vec{\theta}^s)$$
$$\nwarrow p(\vec{\theta}|\mathcal{D}) \simeq \frac{1}{S}\sum_{s=1}^{S}\delta(\vec{\theta}-\vec{\theta}^s)$$



Laplace approximation

Consider $$p(\vec{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\vec{\theta})\,p(\vec{\theta})}{\underbrace{p(\mathcal{D})}_{=Z}} = \frac{p(\vec{\theta},\mathcal{D})}{Z} =$$

$$= \frac{1}{Z} e^{-\mathcal{E}(\vec{\theta})}, \quad \text{where } \mathcal{E} = -\log p(\vec{\theta},\mathcal{D}).$$

Now, expand
$$\mathcal{E}(\vec{\theta}) \simeq \mathcal{E}(\hat{\vec{\theta}}) + (\vec{\theta}-\hat{\vec{\theta}})^T\cdot\vec{g} + \frac{1}{2}(\vec{\theta}-\hat{\vec{\theta}})^T\cdot H\cdot(\vec{\theta}-\hat{\vec{\theta}}),$$
where
$$i,j=1,\dots,D\begin{cases} g_i = \dfrac{\partial\mathcal{E}(\vec{\theta})}{\partial\theta_i}, & \leftarrow \text{gradient} \\[2ex] H_{ij} = \dfrac{\partial^2\mathcal{E}(\vec{\theta})}{\partial\theta_i\,\partial\theta_j} & \leftarrow \text{Hessian} \end{cases}$$

If we expand around the minimum of $\mathcal{E}(\vec{\theta})$, we have $g_i = 0$, $\forall i$ and
$$p(\vec{\theta}|\mathcal{D}) \simeq \frac{1}{Z} e^{-\mathcal{E}(\hat{\vec{\theta}})} e^{-\frac{1}{2}(\vec{\theta}-\hat{\vec{\theta}})^T\cdot \underbrace{H}_{\Sigma^{-1}}\cdot(\vec{\theta}-\hat{\vec{\theta}})}$$

$\int d\vec{\theta}\, p(\vec{\theta}|\mathcal{D}) = 1$ yields
$$\frac{e^{-\mathcal{E}(\hat{\vec{\theta}})}}{Z} = \left[(2\pi)^{D/2}|H|^{-1/2}\right]^{-1}, \text{ or}$$

$$Z = e^{-\mathcal{E}(\hat{\vec{\theta}})}(2\pi)^{D/2}|H|^{-1/2}.$$

Finally, $p(\vec{\theta}|\mathcal{D}) = \mathcal{N}(\vec{\theta}\,|\,\hat{\vec{\theta}}, H^{-1})$.

Bayesian decision theory

Classification: decide the optimal class label to predict, given an observed input $\vec{x}$.

Class labels: $\mathcal{Y} = \{1, \dots, C\}$

If $C = 2$, we can define a zero-one loss

$$\ell_{01}(\underset{\substack{\uparrow \\ \text{observation}}}{y}, \underset{\substack{\uparrow \\ \text{model}}}{\hat{y}}) = \mathbb{I}(y \neq \hat{y})$$

	$\hat{y}=0$	$\hat{y}=1$
$y=0$	0	1
$y=1$	1	0

Then the posterior expected loss

$$R(\hat{y}|\vec{x}) = 1 \times p(\hat{y} \neq y|\vec{x}) + 0 \times p(\hat{y} = y|\vec{x}) =$$

$$= 1 - p(\hat{y} = y|\vec{x}).$$

Thus, to minimize $R(\hat{y}|\vec{x})$ we need to maximize $p(\hat{y} = y|\vec{x})$:

$$\underset{y \in \mathcal{Y}}{\operatorname{argmax}}\, p(y|\vec{x}) \leftarrow$$ choose the most probable value of y (the MAP estimate)

Now, consider the following loss matrix:

	$\hat{y}=0$	$\hat{y}=1$
$y=0$	ℓ_{00}	ℓ_{01}
$y=1$	ℓ_{10}	ℓ_{11}

prob. to pick $\hat{y}=0$ ... prob. to pick $\hat{y}=1$

If $p_0 = p(\hat{y}=0|\vec{x})$ and $p_1 = 1-p_0$,

$$\begin{cases} R(\hat{y}=0|\vec{x}) = \ell_{00}\, p_0 + \ell_{10}\, p_1, \\ R(\hat{y}=1|\vec{x}) = \ell_{01}\, p_0 + \ell_{11}\, p_1 \end{cases}$$

Then we choose $\hat{y}=0$ if

$$\ell_{00}\, p_0 + \ell_{10}\, p_1 < \ell_{01}\, p_0 + \ell_{11}\, p_1 .$$

If $\ell_{00}=0$, $\ell_{11}=0$ (no penalty for correct decisions),

$$\ell_{10}\, p_1 < \ell_{01}(1-p_1), \text{ or}$$

$$p_1 < \frac{\ell_{01}}{\ell_{10}+\ell_{01}}$$

ℓ_{01} = false positive ($\hat{y}=1$ but $y=0$)

ℓ_{10} = false negative ($\hat{y}=0$ but $y=1$)

If, in addition, $\ell_{10} = c\,\ell_{01}$, we pick

$$\hat{y}=0 \text{ iff } p_1 < \frac{1}{1+c}.$$

For ex., if $c=2 \Rightarrow p_1 < \frac{1}{3}$, or $p_0 > \frac{2}{3}$.

So, we declare $\hat{y}=0$ if $p_0 > \frac{2}{3}$, otherwise we declare $\hat{y}=0$.

Note that the 'old' rule (take the $\hat{y}$ with max prob.) corresponds to $c=1$.

Consider thresholding $p(\hat{y}=1|\vec{x})$:

$$p(\hat{y}=1|\vec{x}) \geq 1-\tau \text{ , and define}$$

$$\hat{y}_\tau(\vec{x}) = \mathbb{I}(p(\hat{y}=1|\vec{x}) \geq 1-\tau).$$

Next, for binary data $\{y_n, \vec{x}_n\}_{n=1}^N$ (target: y_n)

define $FP_\tau = \sum_{n=1}^{N} \mathbb{I}(\hat{y}_\tau(\vec{x}_n)=1, y_n=0)$,

and similarly TP_τ, TN_τ, FN_τ.

		Estimate $\hat{y}$: 0	Estimate $\hat{y}$: 1	Row sum
Data y	0	TN_τ	FP_τ	N
	1	FN_τ	TP_τ	P
Column sum		$\hat{N}$	$\hat{P}$	

} normalization for $p(\hat{y}|y)$

$\underbrace{\hat{N} \quad \hat{P}}_{\text{normalization for } p(y|\hat{y})}$

'$p(\hat{y}|y)$ estimates'

		Estimate $\hat{y}$: 0	Estimate $\hat{y}$: 1
Data y	0	$\frac{TN}{N}=TNR$	$\frac{FP}{N}=FPR$
	1	$\frac{FN}{P}=FNR$	$\frac{TP}{P}=TPR$ = recall

R stands for 'rate'

$$\begin{cases} TNR+FPR=1, \\ FNR+TPR=1 \end{cases}$$

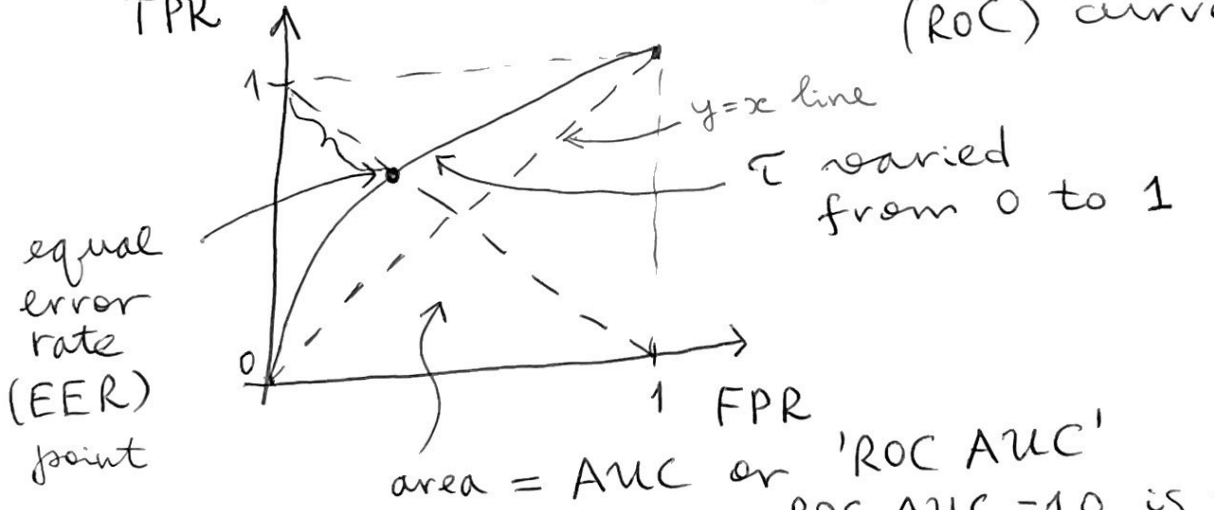
Alternatively, we can focus on 'p(y|$\hat{y}$) estimates':

		Estimate $\hat{y}$ 0	Estimate $\hat{y}$ 1
Data y	0	$\frac{TN}{\hat{N}} = NPV$	$\frac{FP}{\hat{P}} = FDR$
	1	$\frac{FN}{\hat{N}} = FOR$	$\frac{TP}{\hat{P}} = PPV$

NPV = negative predictive value
PPV = pos. pred. value, or precision
FDR = false discovery rate
FOR = false omission rate

Consider
$$\begin{cases} TPR_\tau = p(\hat{y}=1 \mid y=1, \tau) = \dfrac{TP_\tau}{TP_\tau + FN_\tau}, \\ FPR_\tau = p(\hat{y}=1 \mid y=0, \tau) = \dfrac{FP_\tau}{FP_\tau + TN_\tau}. \end{cases}$$

Receiver operating characteristic (ROC) curve



ROC AUC = 1.0 is the best value

EER is defined as FPR = FNR :

$$FPR = 1 - TPR \Rightarrow TPR = 1 - FPR .$$

EER = distance between top left corner and the EER point ; 0 is the best value.

<u>Precision - recall curves</u>

Sometimes, the notion of "negative observation" is not well-defined: e.g. in detecting objects in images, patch by patch, the total number of ~~[illegible]~~ patches is a hyperparameter, so the number of negative observations is user-defined.

In such situations, it may be useful to replace FPR with

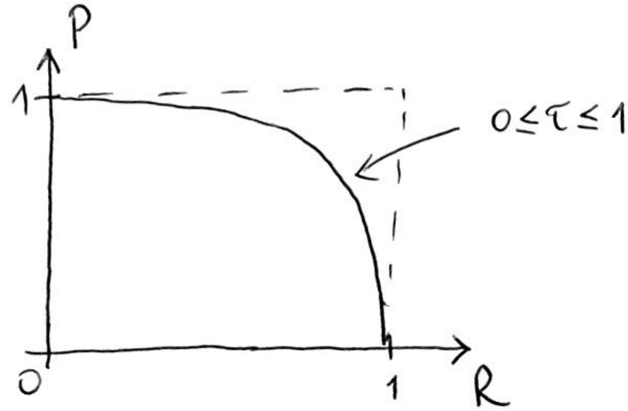
$$\text{precision} = PPV \equiv P(\tau) =$$

$$= p(y=1 \mid \hat{y}=1, \tau) = \frac{TP_\tau}{TP_\tau + FP_\tau} .$$

Likewise,

$$\text{recall} = TPR \equiv R(\tau) = p(\hat{y}=1 \mid y=1, \tau) =$$

$$= \frac{TP_\tau}{TP_\tau + FN_\tau} .$$



The precision-recall curve can be summarized as:

(1) area under the curve (1.0 is the max)

(2) precision of the first K entities recalled "precision at K" score

[Precision-recall curve revisited]

$$\text{Precision} = \frac{TP}{TP + FP}$$

FP = # negative class instances mistakenly identified as positive

$$\text{Recall} = \frac{TP}{TP + FN}$$

FN = # positive class instances mistakenly identified as negative

Thus, $\text{Precision} = \dfrac{\text{Correctly predicted pos.}}{\text{Total predicted pos.}}$

$\text{Recall} = \dfrac{\text{Correctly pred. pos.}}{\text{Truly pos. instances}}$

<u>Ex.</u>

Sports/Not sports

Document	Ground truth	Prediction ($\tau = \frac{1}{2}$, say)	
1	S	S	TP
2	S	S	TP
3	N	S	FP
4	S	N	FN
5	N	N	
6	S	N	FN
7	N	S	FP
8	N	N	

$TP = 2, \quad FP = 2, \quad FN = 2 \implies \begin{cases} Pr = \frac{1}{2} \\ Rc = \frac{1}{2} \end{cases}$

To produce the PR curve, just change the value of the threshold τ.