Kapitel 9

Polynomklassifikator

Ansätze zum Entwurf eines Klassifikators

Es gibt zwei prinzipiell unterschiedliche Ansätze zum Entwurf eines Klassifikators:

1. Statistische parametrische Modellierung der Klassenverteilungen, dann MAP
2. Lösung eines Abbildungsproblems durch Funktionsapproximation (nichtlineare Regression)

$$\min E\left\{ \|\mathbf{f}(\mathbf{x}) - \mathbf{y}\|^2 \right\}$$

$\mathcal{X}$ – *Merkmalsraum*

$\mathcal{Y}$ – *Entscheidungsraum*

Zu 1.) Die bisher beschriebene Vorgehensweise beim Entwurf eines Klassifikators beruht darauf, dass man die *klassenspezifischen Verteilungsdichten*

$$p(\mathbf{x}|\omega)$$

parametrisch durch statistische Modelle annähert (durch Schätzung der Parameter, z.Bsp. Einer Gauß-Verteilung) und durch eine Maximumselektion zu einer Entscheidung kommt. Lernen bedeutet hierbei: Verbesserung des Parameter-Fitting.

Zu 2.) Es gibt nun eine zweite Möglichkeit der Herangehensweise, welche auf die Auswertung der a-posteriori-Wahrscheinlichkeitsdichte

$$p(\omega|\mathbf{x})$$

aufbaut und durch ein Problem der *Funktionsapproximation* beschrieben werden kann.

Diese Funktionsapproximation kann z.Bsp. durch eine *nichtlineare Regression mit Polynomen* oder auch mit Hilfe eines *künstlichen Neuronalen Netzwerkes* (NN) durchgeführt werden. Im Rahmen dieser Veranstaltung sollen die Grundlagen für beide Vorgehensweisen behandelt werden.

Grundsätzlich ist die Suche nach einer besten Näherungsfunktion ein *Variationsproblem*, welches durch die Wahl von Basisfunktionen auf ein *parametrisches Optimierungsproblem* zurückgeführt werden kann. Lernen bedeutet dann auch hier: Parameterfitting.

Die Gleichwertigkeit der beiden genannten Vorgehensweisen ergibt sich aus dem Bayes-Theorem:

$$p(\omega \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega)\,\cancel{p(\omega)}}{\cancel{p(\mathbf{x})}}$$

$p(\omega)$: $1/K$

$p(\mathbf{x})$: unabhängig von ω

Die Gleichwertigkeit ergibt sich daraus, dass der Nenner unabhängig von ω ist und die a-priori-W. i.a. $1/K$ gesetzt werden kann.

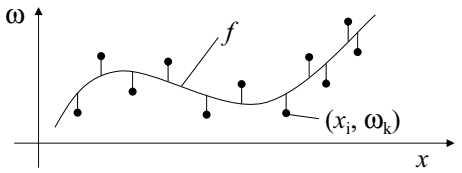
Im folgenden soll nun die Überführung in ein Funktionsapproximationsproblem begründet werden.

Bei bekannter a-posteriori-W. $p(\omega|\mathbf{x})$ würde für jedes kontinuierliche x bestmöglich ein ω zugeordnet werde (funktionale Zuordnung f:$\mathbf{x} \rightarrow \omega$). Gegeben sind jedoch nur *Stichproben* und man sucht nach einer Funktion f, welche bestmöglich die Einzelexperimente fittet und damit eine Abbildung realisert:

$$\mathbf{f}: \underset{\mathbf{x}}{\text{Merkmalsraum}} \rightarrow \underset{\omega}{\text{Bedeutungsraum}}$$

Diese Aufgabenstellung kann mit Hilfe der Variationsrechnung gelöst werden. Wählt man als Gütekriterium das minimale Fehlerquadrat, so geht es um die Minimierung von:

$$J = \min_{\mathbf{f}(\mathbf{x})} E\{\|\mathbf{f}(\mathbf{x}) - \mathbf{y}\|^2\}$$



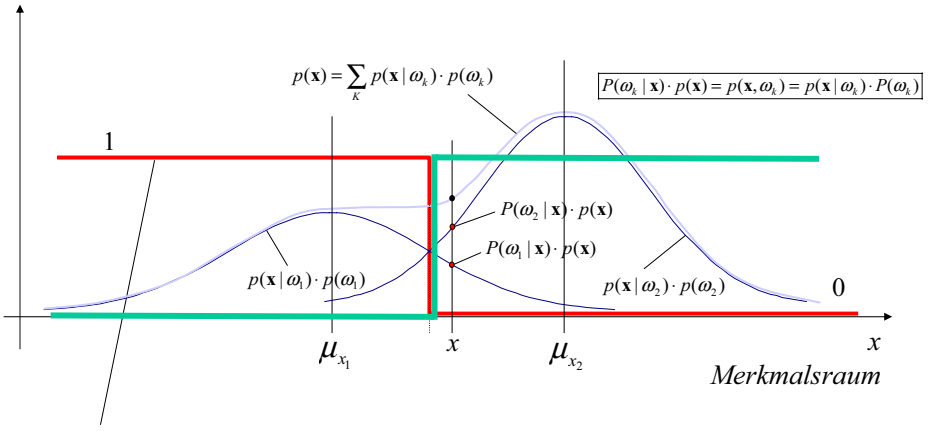
Dabei entstehen die Zielvektoren $\{\mathbf{y}_i\}$ im *Entscheidungsraum* $\mathcal{Y}$ durch einfache Abbildung der skalaren Bedeutungswerte $\{\omega_i\}$

$$\begin{array}{c} \Omega := \{\omega_1, \omega_2, \ldots, \omega_K\} \\ \updownarrow \quad \updownarrow \quad \updownarrow \\ \mathcal{Y} := \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_K\} \end{array}$$

mit:

$$\mathbf{y}_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$ i-ter Einheitsvektor

Zwei-Klassen-Problem mit Gaußverteilungsdichte

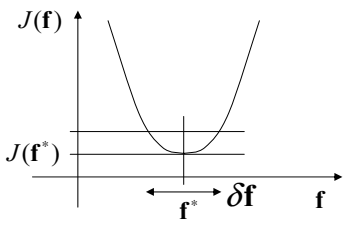


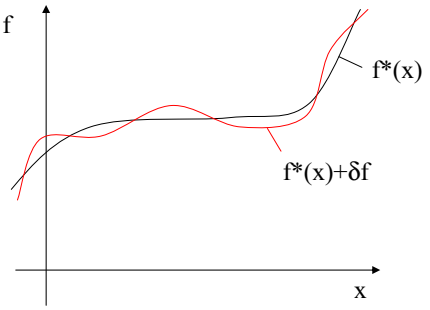
Entscheidungsfunktion für Klasse 1

Lösung des Variationsproblems

Die Lösung des Variationsproblems erhält man über die Annahme, dass man eine optimale Funktion **f*(x)** bereits kennt. Dann muss jede Variation zu einer Verschlechterung des Gütekriteriums führen:

$$J(\mathbf{f}^* + \delta\mathbf{f}) > J(\mathbf{f}^*) \quad \text{für} \quad \forall \delta\mathbf{f} \neq \mathbf{0}$$





Mit Hilfe der Definitionsgleichung ergibt sich:

$$J(\mathbf{f}^* + \delta\mathbf{f}) = E\left\{ \left\| \mathbf{f}^* + \delta\mathbf{f} - \mathbf{y} \right\|^2 \right\} = E\left\{ \langle (\mathbf{f}^* - \mathbf{y}) + \delta\mathbf{f}, (\mathbf{f}^* - \mathbf{y}) + \delta\mathbf{f} \rangle \right\}$$

$$= E\left\{ \left\| \mathbf{f}^* - \mathbf{y} \right\|^2 \right\} + 2E\left\{ \delta\mathbf{f}^T (\mathbf{f}^* - \mathbf{y}) \right\} + E\left\{ \left\| \delta\mathbf{f} \right\|^2 \right\}$$

$$\Rightarrow J(\mathbf{f}^*) = E\left\{ \left\| \mathbf{f}^* - \mathbf{y} \right\|^2 \right\} \quad (\text{für } \delta\mathbf{f} = \mathbf{0})$$

Eingesetzt in die Ungleichung ergibt:

$$E\left\{ \left\| \delta\mathbf{f} \right\|^2 \right\} + 2E\left\{ \delta\mathbf{f}^T (\mathbf{f}^* - \mathbf{y}) \right\} > 0 \qquad \forall \delta\mathbf{f} \neq 0$$

Dies wird erfüllt, wenn der zweite Term verschwindet,
da der erste Term positiv definit ist.

Daraus resultiert eine notwendige Optimalitätsbedingung:

$$E\left\{ \delta\mathbf{f}^T (\mathbf{f}^* - \mathbf{y}) \right\} = 0$$

Zwischenrechnung:

Mit

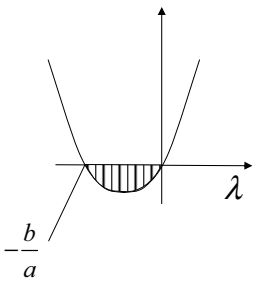
$$a := E\left\{ \left\| \delta\mathbf{f} \right\|^2 \right\} \text{ und } b := 2E\left\{ \delta\mathbf{f}^T (\mathbf{f}^* - \mathbf{y}) \right\}$$

ergibt sich unter der Annahme einer Variation in
eine beliebige Raumrichtung von $\delta\mathbf{f}$:

$$\delta\mathbf{f} \rightarrow \lambda\delta\mathbf{f}$$

$$a\lambda^2 + b\lambda > 0 \quad \forall\lambda$$

$$\lambda(a\lambda + b) > 0 \quad \forall\lambda$$



Ist $b \neq 0$, so gilt für alle $\lambda \in (-b/a, 0)$:

$$\lambda(a\lambda + b) < 0$$

$\Rightarrow$ b muss Null sein, dann lautet die Parabel:

$$a\lambda^2 > 0$$

Der Erwartungswert kann mit der Verbundverteilungsdichte ausformuliert werden zu:

$$E\left\{\delta\mathbf{f}^T(\mathbf{f}^* - \mathbf{y})\right\} = \sum_{\mathbf{x}}\sum_{\mathbf{y}} \delta\mathbf{f}^T(\mathbf{f}^* - \mathbf{y}) \cdot p(\mathbf{x}, \mathbf{y})$$

$$= \sum_{\mathbf{x}} \delta\mathbf{f}^T \left[\sum_{\mathbf{y}} (\mathbf{f}^* - \mathbf{y}) \cdot p(\mathbf{y} \mid \mathbf{x})\right] p(\mathbf{x}) = \mathbf{0}$$

Dieser Ausdruck verschwindet, falls der Term in der eckigen Klammer Null wird.

Daraus resultiert die folgende Optimalitätsbedingung:

$$\sum_{\mathbf{y}} (\mathbf{f}^* - \mathbf{y}) \cdot p(\mathbf{y} \mid \mathbf{x}) = \mathbf{f}^* \sum_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x}) - \sum_{\mathbf{y}} \mathbf{y} \cdot p(\mathbf{y} \mid \mathbf{x})$$

$$= \mathbf{f}^* - \sum_{\mathbf{y}} \mathbf{y} \cdot p(\mathbf{y} \mid \mathbf{x}) = \mathbf{0}$$

Die optimale Schätzfunktion ist die sogenannte Regressionsfunktion:

$$\mathbf{f}^*(\mathbf{x}) = \sum_{\mathbf{y}} \mathbf{y} \cdot p(\mathbf{y} \mid \mathbf{x}) = E\left\{\mathbf{y} \mid \mathbf{x}\right\}$$

Polynomiale Regression

Wahl von Polynomen als Basisfunktionen für die Funktionsapproximation.
Zunächst für eine skalare Funktion mit einem vektoriellen Argument:

$$f(\mathbf{x}) = a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_N x_N$$
$$+ a_{N+1} x_1^2 + a_{N+2} x_1 x_2 + a_{N+3} x_1 x_3 + \ldots$$
$$+ a_{\ldots} x_1^3 + a_{\ldots} x_1^2 x_2 + a_{\ldots} x_1^2 x_3 + \ldots$$

Dieses verallgemeinertes Polynom enthält eine Konstante a_0, gefolgt von N linearen Termen, $N(N+1)/2$ quadratischen Termen usw.

Ein Polynom vom Grade G und der Dimension N des Argumentvektors $\mathbf{x}$ hat

$$L = \binom{N+G}{G} = \frac{(N+G)!}{G!N!}$$

polynomiale Terme, welche als Basisfunktionen $f_i(\mathbf{x})$, i=1,2,...,L einer Funktionsentwicklung betrachtet werden können.

Obiges Polynom kann kompakt durch Einführung einer vektoriellen Abbildung eines *N*-dimensionalen Vektors **x** auf einen *L*-dimensionalen Vektor **p** beschrieben werden:

$$\mathbf{p}(\mathbf{x}) = \begin{bmatrix} 1 & x_1 & x_2 & \dots & x_N & x_1^2 & x_1x_2 & x_1x_3 & \dots & x_1^3 & x_1^2x_2 & x_1^2x_3 & \dots \end{bmatrix}^T$$

Wir führen einen Koeffizientenvektor **a** ein und erhalten für das skalare Polynom:

$$\boxed{f(\mathbf{x}) = \mathbf{a}^T\mathbf{p}(\mathbf{x}) = \sum_{i=0}^{L} a_i p_i(\mathbf{x})}$$

Für die Musterklassifikation benötigen wir für jede Klasse eine Polynomfunktion

$$f_k(\mathbf{x}) = \mathbf{a}_k^T\mathbf{p}(\mathbf{x}) \quad \text{verantwortlich für die Klassen}$$
$$k = 1, 2, \dots, K$$

Kombiniert man die klassenspezifischen Koeffizientenvektoren $\mathbf{a}_k$ zu einer Matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_K \end{bmatrix}$$

so erhält man für die vektorielle Polynomfunktion:

$$\boxed{\mathbf{f}(\mathbf{x}) = \mathbf{A}^T\mathbf{p}(\mathbf{x})}$$

Optimale Anpassung der Matrix **A**

Es gilt nun, die Koeffizientenmatrix A während der Optimierungsprozedur bestmöglich anzupassen.

Dabei kann das Approximationsproblem direkt auf die Messungen oder aber auf die Merkmale eines Unterraumes (z.B. nach einer KLT) angewendet werden.

Adaption der Koeffizientenmatrix **A**:

$$J = \underbrace{\min}_{\mathbf{f}(\mathbf{x})} E\left\{ \left\| \mathbf{f}(\mathbf{x}) - \mathbf{y} \right\|^2 \right\} \approx \underbrace{\min}_{\mathbf{A}} E\left\{ \left\| \mathbf{A}^T\mathbf{p}(\mathbf{x}) - \mathbf{y} \right\|^2 \right\}$$

Das Variationsproblem mit der Suche nach einer unbekannten Funktion wird durch die Wahl polynomialer Basisfunktionen auf ein Parameteroptimierungsproblem reduziert!

Das Gütekriterium *J* minimiert die Varianz des Residuums, nämlich den Fehler zwischen y und der Approximation $\mathbf{A}^T\mathbf{p}(\mathbf{x})$.

Die Lösung wird bestimmt unter Ausnutzung des Variationsansatzes, was zu einer notwendigen Bedingung führt:

$$J(\mathbf{A}^* + \delta\mathbf{A}) \geq J(\mathbf{A}^*) \qquad \text{für} \quad \forall \delta\mathbf{A} \neq \mathbf{0}$$

Es gilt nun, die Koeffizientenmatrix **A** während der Optimierungsprozedur bestmöglich anzupassen.

Adaption der Koeffizientenmatrix **A**:

Wegen

$$J(\mathbf{A}) = E\left\{ \left\| \mathbf{A}^T\mathbf{p}(\mathbf{x}) - \mathbf{y} \right\|^2 \right\} = E\left\{ \left[\mathbf{A}^T\mathbf{p} - \mathbf{y} \right]^T \left[\mathbf{A}^T\mathbf{p} - \mathbf{y} \right] \right\}$$

und $\underbrace{\mathbf{a}^T\mathbf{b}}_{\text{Skalarprodukt}} = \text{Spur}(\underbrace{\mathbf{b}\mathbf{a}^T}_{\substack{\text{dyad.} \\ \text{Produkt}}})$

sowie $\text{Spur}(\mathbf{PQ}) = \text{Spur}(\mathbf{QP})$

erhält man:

$$J(\mathbf{A}) = E\left\{ \text{Spur}\left[\left[\mathbf{A}^T\mathbf{p} - \mathbf{y} \right] \left[\mathbf{A}^T\mathbf{p} - \mathbf{y} \right]^T \right] \right\}$$

$$= \text{Spur}\left[E\left\{ \mathbf{y}\mathbf{y}^T \right\} \right] - 2\,\text{Spur}\left[\mathbf{A}^T E\left\{ \mathbf{p}\mathbf{y}^T \right\} \right] + \text{Spur}\left[\mathbf{A}^T E\left\{ \mathbf{p}\mathbf{p}^T \right\} \mathbf{A} \right]$$

$$= E\left\{ \left\| \mathbf{y} \right\|^2 \right\} - 2\,\text{Spur}\left[\mathbf{A}^T E\left\{ \mathbf{p}\mathbf{y}^T \right\} \right] + \text{Spur}\left[\mathbf{A}^T E\left\{ \mathbf{p}\mathbf{p}^T \right\} \mathbf{A} \right]$$

Wie zuvor angenommen, kann der Zielvektor o.B.d.A. als Einheitsvektor formuliert sein und damit gilt:

$$J(\mathbf{A}) = 1 - 2\,\text{Spur}\left[\mathbf{A}^T E\left\{ \mathbf{p}\mathbf{y}^T \right\} \right] + \text{Spur}\left[\mathbf{A}^T E\left\{ \mathbf{p}\mathbf{p}^T \right\} \mathbf{A} \right]$$

Mit dem Variationsansatz ergibt sich:

$$\begin{aligned} J(\mathbf{A}^* + \delta\mathbf{A}) &= 1 - 2\,\text{Spur}\left[(\mathbf{A}^* + \delta\mathbf{A})^T E\left\{ \mathbf{p}\mathbf{y}^T \right\} \right] \\ &\quad + \text{Spur}\left[(\mathbf{A}^* + \delta\mathbf{A})^T E\left\{ \mathbf{p}\mathbf{p}^T \right\} (\mathbf{A}^* + \delta\mathbf{A}) \right] \\ &= 1 - 2\,\text{Spur}\left[\mathbf{A}^{*T} E\left\{ \mathbf{p}\mathbf{y}^T \right\} \right] - 2\,\text{Spur}\left[\delta\mathbf{A}^T E\left\{ \mathbf{p}\mathbf{y}^T \right\} \right] \\ &\quad + \text{Spur}\left[\mathbf{A}^{*T} E\left\{ \mathbf{p}\mathbf{p}^T \right\} \mathbf{A}^* \right] + 2\,\text{Spur}\left[\delta\mathbf{A}^T E\left\{ \mathbf{p}\mathbf{p}^T \right\} \mathbf{A}^* \right] \\ &\quad + \text{Spur}\left[\delta\mathbf{A}^T E\left\{ \mathbf{p}\mathbf{p}^T \right\} \delta\mathbf{A} \right] \end{aligned}$$

Eingesetzt in die Ungleichung der notwendigen Bedingung für ein Optimum ergibt:

$$\text{Spur}\left[\delta\mathbf{A}^T E\left\{ \mathbf{p}\mathbf{p}^T \right\} \delta\mathbf{A} \right] - 2\,\text{Spur}\left[\delta\mathbf{A}^T \left(E\left\{ \mathbf{p}\mathbf{y}^T \right\} - E\left\{ \mathbf{p}\mathbf{p}^T \right\} \mathbf{A}^* \right) \right] \geq 0$$

Da der erste Term positiv definit ist und wegen der Positiv-Definitheit von der Momentenmatrix $E\{\mathbf{p}\mathbf{p}^T\}$, muss der zweite Term verschwinden:

$$\text{Spur}\left[\delta\mathbf{A}^T\left(E\left\{\mathbf{p}\mathbf{y}^T\right\}-E\left\{\mathbf{p}\mathbf{p}^T\right\}\mathbf{A}^*\right)\right]=0 \qquad \forall\delta\mathbf{A}$$

und daraus schließlich die in **A** lineare Gleichung:

$$\boxed{E\left\{\mathbf{p}\mathbf{p}^T\right\}\mathbf{A}^*=E\left\{\mathbf{p}\mathbf{y}^T\right\}}$$

diese Gleichung garantiert ein minimales Residuum des Fehlervektors:

$$\Delta\mathbf{f}(\mathbf{x})=\mathbf{f}(\mathbf{x})-\mathbf{y}=\mathbf{A}^T\mathbf{p}(\mathbf{x})-\mathbf{y}$$

Die Lösung des Problems führt auf die Aufgabe, die beiden Momentenmatrizen $E\{\mathbf{p}\mathbf{p}^T\}$ und $E\{\mathbf{p}\mathbf{y}^T\}$ aus dem Trainingsdatensatz zu schätzen und eine lineare Matrizengleichung zu lösen. Der verbleibende Fehlervektor berechnet sich zu:

$$\boxed{J(\mathbf{A}^*)=1-\text{Spur}\left[\mathbf{A}^{*T}E\left\{\mathbf{p}\mathbf{p}^T\right\}\mathbf{A}^*\right]}$$

Orthogonalität des Schätzfehlervektors

Die Bestapproximation fordert nach dem Projektionssatz, dass der Fehlervektor $\Delta\mathbf{f}$ senkrecht steht auf den Unterraum, welcher durch die polynomialen Basisfunkionen in $\mathbf{p}(\mathbf{x})$ aufgespannt wird.

In statistischer Notation bedeutet das, dass die Momentenmatrix, welche aus **p** und $\Delta\mathbf{f}$ aufgespannt wird, eine Nullmatrix sein muss:

$$E\left\{\mathbf{p}\Delta\mathbf{f}^T\right\}=E\left\{\mathbf{p}\left(\mathbf{A}^{*T}\mathbf{p}-\mathbf{y}\right)^T\right\}=E\left\{\mathbf{p}\mathbf{p}^T\right\}\mathbf{A}^*-E\left\{\mathbf{p}\mathbf{y}^T\right\}=0$$

Diese Eigenschaft wird auch von **p** auch auf **f** vererbt, da **f** aus Linearkombinationen der Basisfunktionen in $\mathbf{p}(\mathbf{x})$ aufgebaut wird.

$$E\left\{\mathbf{f}^*\Delta\mathbf{f}^T\right\}=E\left\{\mathbf{A}^{*T}\mathbf{p}\Delta\mathbf{f}^T\right\}=\mathbf{A}^{*T}E\left\{\mathbf{p}\Delta\mathbf{f}^T\right\}=\mathbf{0}$$

Erwartungstreue der Approximationsfunktion

Die Schätzfunktion ist erwartungstreu (unbiased). Der Schätzfehler (Residuum) hat den Erwartungswert **0**.

$$E\{\Delta\mathbf{f}\} = E\{\mathbf{f}(\mathbf{x}) - \mathbf{y}\} = \mathbf{0}$$

Daraus folgt, dass **f**(**x**) eine erwartungstreue Schätzung von **y** ist:

$$E\{\mathbf{f}(\mathbf{x})\} = E\{\mathbf{y}\}$$

Rekursive Lernregel für den Polynomklassifikator

Für den auf das minimale Fehlerquadrat aufbauende Klassifikator ergibt sich die folgende rekursive Lernstrategie für die Koeffizientenmatrix **A**:

$$\mathbf{A}_n = \mathbf{A}_{n-1} - \alpha\left[\sum_{i=1}^{n}\mathbf{p}_i\mathbf{p}_i^T\right]^{-1}\mathbf{p}_n\left[\mathbf{A}_{n-1}^T\mathbf{p}_n - \mathbf{y}_n\right]^T$$

Diese Lernregel beinhaltet folgende Schritte:

- Wahl eines Startwerts für **A**
- Die momentante Stichprobe $[\mathbf{x},\mathbf{y}]$ wird abgebildet mit Hilfe der Polynombasis **p**(**x**) auf das Paar $[\mathbf{p},\mathbf{y}]$
- Berechne die Schätzung **f** auf der Basis der gegebenen Beobachtung **p** und der momentanten Koeffizientenmatrix **A**: $\mathbf{f} = \mathbf{A}^T\mathbf{p}$
- Berechne das Residuum: $\Delta\mathbf{f} = \mathbf{f} - \mathbf{y}$
- Korrigiere $\mathbf{A}_n$ auf der Grundlage obiger Rekursion

$$\mathbf{A}_n = \mathbf{A}_{n-1} - \alpha\mathbf{G}^{-1}\mathbf{p}_n\Delta\mathbf{f}^T$$

Die Gewichtsmatrix **G** ist dabei die Momentenmatrix $E\{\mathbf{p}\mathbf{p}^T\}$ auf der Basis der n zur Verfügung stehenden Stichproben.

Eine strikte Beachtung obiger Rekursion, erfordert eine weitere Rekursion für die Inverse der Gewichtsmatrix gemäß:

$$\mathbf{G}_n^{-1} = \tfrac{1}{1-\alpha}\left[\mathbf{G}_{n-1}^{-1} - \alpha\frac{\mathbf{G}_{n-1}^{-1}\mathbf{p}_n\mathbf{p}_n^T\mathbf{G}_{n-1}^{-1}}{1+\alpha\left(\mathbf{p}_n^T\mathbf{G}_{n-1}^{-1}\mathbf{p}_n - 1\right)}\right]$$

G oder die echte zusätzliche Iteration ändern fast nichts am Ergebnis. G kann sogar noch vereinfacht werden zu:

$$\mathbf{G} = E\left\{\mathbf{p}\mathbf{p}^T\right\} = \sum_{i=1}^{n}\mathbf{p}_i\mathbf{p}_i^T$$

Herleitung der rekursiven Lernregel

$$E\left\{\mathbf{p}\mathbf{p}^T\right\}\mathbf{A} = E\left\{\mathbf{p}\mathbf{y}^T\right\}$$

Einführung von Schätzwerten:

$$\underbrace{\left(\tfrac{1}{n}\sum_{i=1}^{n}\mathbf{p}_i\mathbf{p}_i^T\right)}_{\mathbf{G}_n}\mathbf{A}_n = \underbrace{\tfrac{1}{n}\sum_{i=1}^{n}\mathbf{p}_i\mathbf{y}_i^T}_{\mathbf{H}_n}$$

$\mathbf{G}_n\mathbf{A}_n = \mathbf{H}_n$ und $\mathbf{G}_n = (1-\alpha)\mathbf{G}_{n-1} + \alpha\mathbf{p}_n\mathbf{p}_n^T$, $\mathbf{H}_n = (1-\alpha)\mathbf{H}_{n-1} + \alpha\mathbf{p}_n\mathbf{y}_n^T$

$$\Rightarrow \mathbf{G}_n\mathbf{A}_n = (1-\alpha)\mathbf{H}_{n-1} + \alpha\mathbf{p}_n\mathbf{y}_n^T$$
$$= (1-\alpha)\mathbf{G}_{n-1}\mathbf{A}_{n-1} + \alpha\mathbf{p}_n\mathbf{y}_n^T$$
$$= \left(\mathbf{G}_n - \alpha\mathbf{p}_n\mathbf{p}_n^T\right)\mathbf{A}_{n-1} + \alpha\mathbf{p}_n\mathbf{y}_n^T$$
$$= \mathbf{G}_n\mathbf{A}_{n-1} - \alpha\mathbf{p}_n(\mathbf{p}_n^T\mathbf{A}_{n-1} - \mathbf{y}_n^T)$$
$$\Rightarrow \boxed{\mathbf{A}_n = \mathbf{A}_{n-1} - \alpha\mathbf{G}_n^{-1}\mathbf{p}_n\left[\mathbf{A}_{n-1}^T\mathbf{p}_n - \mathbf{y}_n\right]^T}$$

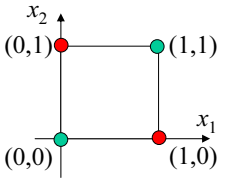
Polynomklassifikator für das XOR-Problem

Wir wählen als Regressionsfunktion ein Polynom mit Termen zweiten Grades in der Hoffnung, dass damit eine Lösung erreicht werden kann:

$$f(\mathbf{x}) = a_1 x_1 + a_2 x_2 + a_3 x_1 x_2 = \mathbf{a}^T \mathbf{p}(\mathbf{x})$$

mit:

$$\mathbf{p} = \left[x_1, x_2, x_1 x_2\right]^T \quad \text{und} \quad \mathbf{a} = \left[a_1, a_2, a_3\right]^T$$

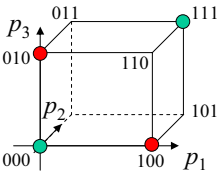


$\mathcal{X}$-Raum

Die Eckpunkte des Quadrats im zweidimensionalen Ursprungsraum $\mathcal{X}$ werden auf die Ecken eines (Hyper-) Würfels einen dreidimensionalen Merkmalsraum $\mathcal{P}$ abgebildet und dieser schließlich auf den eindimensionalen Bedeutungsraum $\mathcal{Y}$:

$$\mathcal{X} \rightarrow \mathcal{P} \rightarrow \mathcal{Y}$$

$\mathcal{X}$	x_1	0	1	0	1
	x_2	0	0	1	1
	p_1	0	1	0	1
$\mathcal{P}$	p_2	0	0	1	1
	p_3	0	0	0	1
$\mathcal{Y}$	y	0	1	1	0



$\mathcal{P}$-Raum

Die optimalen Parameter der Regressionsfunktion **a*** ergeben sich aus:

$$\underbrace{E\left\{\mathbf{p}\mathbf{p}^T\right\}}_{\mathbf{R}_{pp}}\mathbf{a}^* = \underbrace{E\left\{\mathbf{p}\mathbf{y}^T\right\}}_{\mathbf{R}_{py}} \quad \text{bzw:} \quad \mathbf{a}^* = \mathbf{R}_{pp}^{-1} \cdot \mathbf{R}_{py}$$

mit: $\mathbf{R}_{pp} = E\left\{\mathbf{p}\mathbf{p}^T\right\} \approx \frac{1}{4}\sum_{i=1}^{4} \mathbf{p}_i \mathbf{p}_i^T$

$$= \frac{1}{4}\left\{\mathbf{0} + \left[\begin{array}{c|ccc} & 1 & 0 & 0 \\ \hline 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] + \left[\begin{array}{c|ccc} & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] + \left[\begin{array}{c|ccc} & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{array}\right]\right\} = \frac{1}{4}\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

und: $\mathbf{R}_{py} = E\left\{\mathbf{p}y\right\} \approx \frac{1}{4}\sum_{i=1}^{4} \mathbf{p}_i y_i = \frac{1}{4}\sum_{i=1}^{4} y_i \mathbf{p}_i$

$$= \frac{1}{4}\left\{0 \cdot \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + 1 \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 1 \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 0 \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}\right\} = \frac{1}{4}\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

Daraus folgt:

$$\mathbf{a}^* = \mathbf{R}_{pp}^{-1} \cdot \mathbf{R}_{py} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 3 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$$

und damit eine Regressionsfunktion

$$\boxed{f(\mathbf{x}) = a_1 x_1 + a_2 x_2 + a_3 x_1 x_2 = x_1 + x_2 - 2x_1 x_2}$$

welche an den 4 Funktionswerten genau die Werte der Zielfunktion annimmt und ansonsten interpoliert!